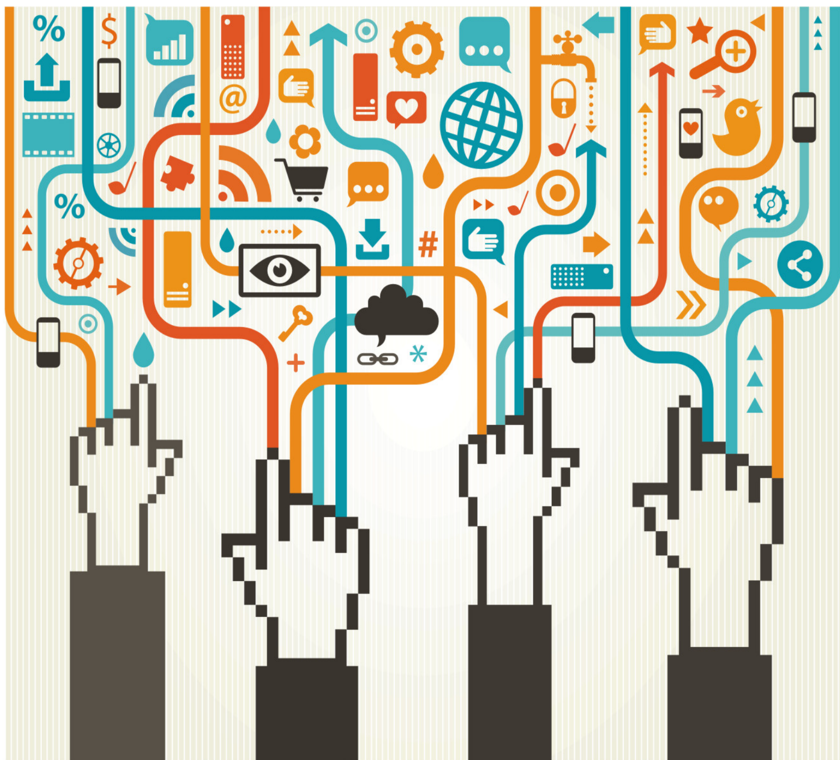
Sociologie
Antropologie

Marian Vasile

INTRODUCERE ÎN SPSS

pentru cercetarea socială şi de piaţă



Collegium

POLIROM

Collegium. Sociologie. Antropologie – serie coordonată de Lazăr Vlăsceanu
şi Liviu Chelcea şi iniţiată de Elisabeta Stănciulescu.

www.polirom.ro

Editura POLIROM
Iaşi, B-dul Carol I nr. 4; P.O. BOX 266, 700506
Bucureşti, Splaiul Unirii nr. 6, bl. B3A, sc. 1, et. 1,
sector 4, 040031, O.P. 53, C.P. 15-728

Descrierea CIP a Bibliotecii Naţionale a României:

VASILE, MARIAN

Introducere în SPSS pentru cercetarea socială şi de piaţă: o perspectivă aplicată / Marian Vasile. – Iaşi: Polirom, 2014
Bibliogr.

ISBN print: 978-973-46-4136-9
ISBN ePub: 978-973-46-4460-5
ISBN PDF: 978-973-46-4462-9

004.42 SPSS:339.13+316.65
519.22:339.13+316.65

Printed in ROMANIA

Marian Vasile

INTRODUCERE ÎN SPSS

pentru cercetarea socială şi de piaţă

O perspectivă aplicată

POLIROM
2014

MARIAN VASILE este lector universitar la Universitatea din București, Facultatea de Sociologie și Asistență Socială, și cercetător la Academia Română, Institutul de Cercetare a Calității Vieții. A încheiat un proiect postdoctoral centrat pe analizarea mecanismelor care influențează fericirea indivizilor. Este doctor în sociologie din 2010, cu o teză în care identifică o serie de stiluri de viață specifice României anilor 2000. A participat la numeroase cursuri despre regresia multinivel și modele de ecuații structurale şi la proiecte de cercetare socială în domenii precum stilurile de viață, stratificarea socială, educația, calitatea vieții şi valorile. Este membru al European Sociological Association și International Sociological Association. Mai multe informații pe http://www.marian-vasile.ro.

Cuprins

Lista figurilor şi tabelelor

1. Introducere

1.1. Cui este utilă această carte

De câţiva ani buni, îi ajut pe studenţi să înţeleagă utilitatea statisticii în cercetarea socială şi de marketing. Implicit, pentru că realizarea analizelor statistice fără un software dedicat este dificil de imaginat astăzi, încerc să îi familiarizez cu unul dintre acestea. Pentru că studenţii cu care lucrez sunt, într-un număr destul de mare, absolvenţi de filologie sau de ştiinţe sociale, acest demers este o provocare, una plăcută însă. Programul de statistică utilizat în această lucrare este **IBM® SPSS® Statistics software (SPSS)**[1], versiunea 17. Toate operaţiunile pot fi reproduse folosind orice versiune recentă a programului. Vă recomand versiunile mai noi, pentru că sintaxa reliefează, folosind culori, diferitele elemente care o compun. Va fi mai uşor să vă obişnuiţi cu aceasta.

Studenţii optează pentru un curs doar dacă acesta li se pare util. Statistica este cât se poate de utilă în orice domeniu, dar în ştiinţele sociale este destul de greu să îi convingi pe cei care se tem de matematică să aleagă de bunăvoie şi nesiliţi de nimeni să treacă prin acest „calvar”. Unii studenţi procedează în felul următor : deschid progamul de statistică, în cazul de faţă SPSS, şi încearcă să reproducă paşii explicaţi în diferite manuale sau tutoriale. Inevitabil, interacţionează cu concepte din statistică, dar le acordă mai puţină importanţă în procesul de învăţare decât meniurilor şi comenzilor din program. Aceasta este o perspectivă „inversă”. Nu poţi învăţa să foloseşti un program de statistică dacă nu ştii… statistică. Este ca şi când ai vrea să devii pilot de Formula 1 fără să ai permis de conducere. Această abordare duce la învăţare mecanică : utilizatorul intră în meniurile SPSS şi dă clickuri ici şi colo fără să-i fie clar de ce face aceste lucruri, de ce alege o opţiune, şi nu alta, sau cum sunt interpretate rezultatele pe care le oferă aceste acţiuni.

Cum ar trebui să procedeze studentul ? Ar trebui să parcurgă un manual de statistică şi, simultan, un manual în care analizele statistice sunt puse în practică într-un program de statistică – SPSS. Cursurile de statistică teoretică sau aplicată nu sunt niciodată suficiente. Domeniul este atât de dezvoltat, încât subiectul nu poate fi epuizat într-o singură lucrare. Învăţarea statisticii este un proces. Cel sau cea care se angajează în acest demers trebuie să adauge consultării materialelor teoretice multe exerciţii folosind date reale. Astăzi este foarte la îndemână acest lucru. Tot mai

1. SPSS Inc. a fost achiziţionat de IBM în octombrie 2009.

multe date sunt accesibile gratuit. Vedeţi în acest sens studiile European Values Study, World Values Survey, European Social Survey, Eurobarometrul etc. Pe paginile web ale acestor cercetări găsiţi chestionarele utilizate, documentaţie extensivă despre activitatea de teren, baze de date şi multe alte informaţii care vă ajută să înţelegeţi complexitatea abordării cantitative a realităţii şi tipul de rezultate care pot fi obţinute astfel. Să presupunem că Maria şi-a dat seama că statistica este importantă şi s-a decis să înveţe principalele tehnici utilizate în piaţă. Dar este abia în anul I de facultate, astfel că nu a avut ocazia să participe la cercetări în calitate de analist. Adică nu i-a pus nimeni în braţe un chestionar, o bază de date şi o listă de întrebări de cercetare pentru soluţionarea cărora să fie nevoită să facă anumite analize statistice. În această situaţie, ar putea să rezolve exerciţiile din manualele de statistică folosind, evident, programul SPSS. Din experienţa proprie, pot spune că, într-un final, va ajunge să înţeleagă multe lucruri, dar pe parcurs s-ar putea să se descurajeze şi să aibă impresia că drumul pe care s-a angajat este foarte greu şi nu tocmai plăcut. Dacă nu este autodidactă sau foarte hotărâtă, atunci Maria s-ar putea să renunţe la un moment dat.

Consider că lipseşte un manual care să îl ajute pe studentul începător în cercetare să unească logica activităţii de cercetare în ştiinţele sociale şi logica manualelor de statistică. Principalele întrebări la care răspunde acest volum sunt :

- Care este legătura dintre chestionarul care a fost utilizat pentru a culege date şi baza de date ?
- Cum realizaţi o bază de date ?
- Ce înseamnă să curăţaţi baza de date ?
- Ce înseamnă să pregătiţi datele pentru analiză ?
- Ce sunt codificarea şi recodificarea unei variabile ?
- Cum creaţi variabile într-o bază de date ?
- De ce trebuie să vă uitaţi la date, înainte de a face analiza care vă interesează ?
- Cum faceţi această explorare primară a datelor ?
- Ce este un tabel ? Dar un tabel de contingenţă ?
- Cum verificaţi dacă variabila X este asociată cu variabila Y ?
- Care este diferenţa dintre asociere şi corelaţie ?
- Dacă doriţi să explicaţi un fenomen, să zicem fericirea (Y), iar teoriile vă spun că este posibil ca acesta să fie explicat de mai mulţi factori, să zicem starea de sănătate (X1), calitatea relaţiilor sociale (X2) şi cantitatea de timp liber avută la dispoziţie (X3), ce tehnică statistică puteţi folosi în acest sens ?

Lista nu este completă. Pe măsură ce citiţi acest volum, puteţi adăuga întrebările la care aţi găsit un răspuns. Statistica oferă mai multe metode prin care putem răspunde la aceeaşi întrebare. SPSS oferă mai multe comenzi pentru aceeaşi analiză. Le voi prezenta pe cele mai importante pentru cei aflaţi la început de drum. Tranziţia spre lucrurile mai dificile va fi mai uşoară după ce aţi parcurs acest volum.

Introducere în SPSS pentru cercetarea socială şi de piaţă. O perspectivă aplicată se adresează, în primul rând, studenţilor care vor să facă primii paşi în abordarea cantitativă a socialului. Ei pot fi studenţi la sociologie, marketing sau

administrarea afacerilor. Logica este în multe situaţii similară în aceste domenii. Apoi, sunt vizaţi masteranzii care au o pregătire limitată în statistică şi utilizarea SPSS-ului, dar şi doctoranzii care nu au urmat un curs intensiv în acest domeniu şi nici nu au lucrat în multe proiecte care folosesc date cantitative. De asemenea, cred că este util şi pentru cercetătorii care au utilizat SPSS, dar l-au învăţat, mai degrabă, „pe încercate”, şi nu în mod sistematic.

1.2. Ce conţine această carte şi cum să o citim

Acord o atenţie considerabilă aspectelor premergătoare activităţii de analiză cantitativă a datelor culese prin aplicarea unor chestionare. O mare parte din timpul activităţii de analiză este consumat de aceste aspecte preliminare. Primele elemente care îl preocupă pe cercetătorul cantitativist sunt elaborarea bazei de date (capitolul 2) şi curăţarea acesteia (capitolul 4). Pentru procesul de curăţare, acesta trebuie să înveţe câteva operaţiuni cum ar fi filtrarea bazei de date (capitolul 3) sau crearea de variabile noi (capitolul 5). Este dificil să scrii o lucrare care urmăreşte toţi aceşti paşi, exact în ordinea în care se întâmplă în realitate. Demersul este circular, de aceea, de exemplu, în procesul de curăţare voi folosi informaţii prezentate şi în capitolele ulterioare, cum ar fi cele despre tabelele de contingenţă (capitolul 6). Cert este că informaţiile din capitolele 2, 3, 4 şi 5 sunt esenţiale şi trebuie citite înainte de a trece la capitolul 6. Odată cu capitolul 6, cititorul primeşte şi informaţii despre analizele statistice uzuale care pot fi utilizate pentru a răspunde la întrebări de cercetare. Cum observăm modul în care gândeşte majoritatea? Cât de omogene sunt diferite grupuri în funcţie de o anumită caracteristică? Media, mediana, abaterea standard şi altele sunt doar câteva elemente utile pentru a răspunde la astfel de întrebări. Tabelul de contingenţă ne va ajuta să vedem dacă două variabile sunt independente sau nu. Apoi, aflăm cum putem testa diferenţa dintre două sau mai multe grupuri în funcţie de o caracteristică. După aceea, aflăm cum explicăm variaţia unei variabile în funcţie de mai multe caracteristici. Media generală la învăţătură a elevilor care au făcut trei ore de educaţie fizică pe săptămână la şcoală este mai ridicată decât cea a elevilor care au făcut cel mult o oră de educaţie fizică pe săptămână la şcoală? Volumul vânzărilor iaurtului cu căpşuni produs de firma „Iaurt pentru toţi” este mai mare dacă în hipermarketuri se foloseşte testarea produsului (adică firma a angajat promotori care le oferă potenţialilor cumpărători să guste iaurtul respectiv) decât dacă nu se foloseşte? Informaţiile prezentate în acest volum pot fi utilizate atât în situaţii întâlnite în cercetarea socială, cât şi în cea de piaţă. La acest gen de întrebări putem răspunde statistic folosind informaţiile din capitolul 6. Realitatea socială este mult mai complexă. Nu ne putem aştepta ca media generală la învăţătură a elevilor să depindă doar de practicarea frecventă a unor activităţi sportive, la fel cum nu ne putem aştepta ca volumul vânzărilor unui tip de iaurt

să depindă doar de prezenţa promotorilor în magazine. Am putea adăuga, pentru primul caz, numărul de ore petrecute în bibliotecă studiind individual, ajutorul primit din partea părinţilor, participarea la activităţi extraşcolare cu caracter educativ, numărul colegilor sau prietenilor cu care elevul îşi petrece timpul liber, caracteristicile acestora etc. În al doilea caz, am putea adăuga calitatea distribuţiei, atractivitatea ambalajului, preţul produsului, poziţionarea la raft etc. Avem, aşadar, o variabilă dependentă şi mai multe variabile independente. Pentru acest gen de situaţii, informaţiile prezentate în capitolul 8 vor fi utile.

Capitolul 7 tratează o serie de asumpţii fundamentale pentru analizele statistice prezentate în capitolele 6 şi 8. Aş fi putut opta pentru o prezentare sumară în cadrul fiecărui capitol, însă am vrut să subliniez importanţa acestui pas. Domeniul explorării asumpţiilor este vast, depăşind obiectivele acestui volum care constituie, în primul rând, un material introductiv. Pe măsură ce învăţaţi mai multe analize statistice, în special multivariate, veţi identifica şi alte asumpţii care trebuie testate.

Toate capitolele se încheie cu o listă de exerciţii care pot fi folosite pentru a pune în practică informaţiile prezentate pe parcursul capitolului respectiv. Exerciţiile înseamnă experienţă acumulată atât cu conceptele, cât şi cu programul de statistică. Consider că niciodată nu facem suficiente exerciţii, aşadar lista cu exerciţii de la finalul fiecărui capitol este doar un prolog al eforturilor dumneavoastră viitoare.

1.3. Materiale suplimentare

Puteţi descărca date utilizate pentru diferite exemple, sintaxe şi outputuri produse de aceste sintaxe de pe pagina de internet: http://www.marian-vasile.ro/publications/spss.

1.4. Mulţumiri

Aş vrea să le mulţumesc celor de la care am învăţat, la rândul meu, multe dintre lucrurile pe care le ştiu atât despre analizele statistice, cât şi despre utilizarea SPSS. În primul rând, vreau să îi mulţumesc lui Bogdan Voicu, care nu numai că mi-a răspuns la toate întrebările, dar mi-a oferit şi oportunitatea de a-i fi asistent la cursurile sau trainingurile ţinute în diferite contexte. Apoi, aş vrea să îi mulţumesc lui Alexandru Cernat pentru că a acordat timp citirii acestui material, oferindu-mi sugestii extrem de utile. Îi mulţumesc lui Ioan Mărginean pentru că m-a provocat să gândesc critic diferite situaţii întâlnite în cercetarea calităţii vieţii, şi nu numai. Nu în ultimul rând, le mulţumesc lui Liviu Chelcea şi Lazăr Vlăsceanu pentru că m-au încurajat să public această lucrare.

Multe dintre informaţiile acumulate şi transpuse, într-o formă sau alta, în acest volum au fost acumulate în cadrul unor proiecte de cercetare similare cu cel postdoctoral susţinut de UEFISCDI, care s-a derulat între 2011 şi 2013 sub denumirea *Drumuri diferite către* o *viaţă mai bună* : *comparaţii internaţionale longitudinale ale determinanţilor satisfacţiei cu viaţa* (PN-II-RU-PD-2011-3-0117). Un alt proiect este cel coordonat de Bogdan Voicu, care s-a derulat între 2011 şi 2014 sub titlul *Schimbarea socială în contextul migraţiei internaţionale : patternuri valorice, participare civică şi politică, satisfacţia cu viaţa* (PN-II-ID-PCE-2011-3-0210). Pentru mai multe detalii, puteţi consulta paginile de internet http : //www.stilurideviata.ro şi http : //www.romanianvalues.ro.

2. Crearea unei baze de date

„Cercetare cantitativă” sau „analiză cantitativă” sunt două concepte frecvent folosite de practicieni în activitatea de zi cu zi. Ambele fac trimitere la culegerea şi analiza unor informaţii prin utilizarea chestionarului ca instrument de cercetare. Chestionarul cuprinde o serie de întrebări închise şi, uneori, şi întrebări deschise. Întrebările închise au răspunsurile predefinite de cercetător, persoana care este rugată să răspundă la întrebare (respondentul) trebuind doar să îl aleagă pe cel care i se potriveşte cel mai bine. Întrebările deschise nu au răspunsuri predefinite, respondentul trebuind să compună, folosind cuvintele proprii, un răspuns care caracterizează cel mai bine modul cum gândeşte, se comportă sau, mai general, care prezintă situaţia sa la momentul intervievării sau la cel de referinţă folosit de cercetător. Tabelul 2.1 prezintă un exemplu care diferenţiază aceste două tipuri de întrebări.

Tabelul 2.1. Două tipuri de întrebări folosite în chestionare : închise şi deschise

Întrebare închisă	**Întrebare deschisă**
D2. Ocupaţia dvs. actuală (principală) : 1. agricultor 2. muncitor (meseriaş) 3. tehnician, maistru, funcţionar 4. ocupaţii cu studii superioare 5. altă ocupaţie 6. elev, student 7. pensionar 8. casnică 9. acum sunt şomer 10. patron	Q112. Cum se numeşte munca pe care o desfăşuraţi (la principalul loc de muncă) ? ______________________________ Q112a. Ce fel de activitate desfăşuraţi în cea mai mare parte a timpului ? ______________________________
Sursa : *Diagnoza calităţii vieţii din România*, Institutul de Cercetare a Calităţii Vieţii, 2010.	*Sursa* : *European Values Study*, Institutul de Cercetare a Calităţii Vieţii, 2008.

În exemplul din tabelul 2.1, cercetătorul este interesat să afle structura ocupaţiilor din România. Dacă foloseşte întrebarea închisă, atunci respondentul va alege varianta de răspuns care se potriveşte cel mai bine situaţiei sale. Dacă foloseşte întrebarea deschisă, atunci respondentul va descrie în cuvinte, cât mai detaliat, situaţia sa cu privire la acest subiect. De regulă, preferăm să folosim întrebări închise în chestionare pentru că aplicarea acestora durează mai puţin, sunt mai

uşor de înţeles, se introduc mai repede în baza de date, ne reprezentăm mai uşor ce fel de analize statistice putem realiza cu ele etc. Dacă optează pentru această variantă, cercetătorul trebuie să se asigure că lista variantelor de răspuns este completă, iar acestea nu se suprapun, adică respondentul nu se poate regăsi în mai multe răspunsuri simultan. Există situaţii, însă, în care mai multe răspunsuri sunt plauzibile pentru aceeaşi persoană, acestea fiind surprinse prin întrebările cu răspuns multiplu. Dacă optează pentru varianta cu întrebări deschise, atunci cercetătorul trebuie să ştie cum va codifica răspunsurile primite. Codificarea presupune ca, din lista lungă de răspunsuri primite, să construiască una mai restrânsă, astfel încât fiecare categorie să poată primi un cod unic care va fi introdus în baza de date şi, ulterior, va fi folosit pentru diferite analize statistice. De exemplu, întrebarea D2 (tabelul 2.1), are coduri de la 1 la 10. În acest exemplu, codificarea la Q112 şi Q112a (tabelul 2.1), va fi realizată folosind o schemă de coduri standardizată, *International Standard Classification of Occupations*[1] (ISCO ; Clasificarea internaţională standard a ocupaţiilor). Aceasta are mai multe variante. Dacă ne uităm la ISCO-88, putem vedea că, la nivelul cel mai înalt de generalitate, din răspunsurile deschise putem obţine zece coduri, numerotate de la 0 la 9. Fiecare dintre aceste categorii ocupaţionale largi poate fi divizată în mai multe subgrupuri. La cel mai rafinat nivel de specificare se poate ajunge la 390 de grupuri ocupaţionale, adică 390 de coduri. Nivelul de detaliu ales de cercetător depinde, în mare măsură, de volumul eşantionului pe care îl foloseşte.

Informaţii despre tipurile de întrebări, regulile de elaborare a acestora, opţiunea pentru o formă sau alta şi nu numai pot fi găsite în lucrările dedicate subiectului cum ar fi cele elaborate de Mărginean (1982), Bradburn, Sudman *et al*. (2004), Chelcea (2007), Saris şi Gallhofer (2007) sau Malhotra (2007). Acestea sunt cunoştinţe complementare celor prezentate aici şi trebuie însuşite pentru o înţelegere adecvată a procesului cercetării cantitative.

După aplicarea chestionarelor, acestea trebuie introduse în baza de date. Apoi baza de date trebuie curăţată. Abia după aceste etape, putem trece la analizele statistice prin care răspundem la întrebările de cercetare. În acest capitol vom afla cum se realizează o bază de date în care sunt introduse chestionare şi cum ajungem la baza de date în format SPSS. În capitolul 3 vom învăţa câteva comenzi esenţiale pentru gestionarea bazei de date, iar în capitolul 4 vom afla care sunt etapele procesului de curăţare a bazei de date şi ce presupune fiecare dintre ele.

Să presupunem că avem 1.000 de chestionare care trebuie introduse într-o bază de date. Pentru aceasta, există mai multe opţiuni. Astăzi, din ce în ce mai multe institute şi companii de cercetare socială şi/sau de piaţă înlocuiesc chestionarele pe hârtie cu chestionarele în format digital. Adică operatorul de teren nu

1. Documentele despre această clasificare pot fi consultate pe pagina dedicată de pe site-ul ISCO : http : //www.ilo.org/public/english/bureau/stat/isco/isco88/publ3.htm.

mai completează cu pixul pe hârtie răspunsurile la întrebări, ci dă click pe un laptop sau pe o tabletă. Formatul digital de aplicare a chestionarelor are mai multe avantaje faţă de cel clasic, pe hârtie. Baza de date, chiar şi cea în format SPSS, este creată direct, fiind redus astfel necesarul de resurse umane, timp şi bani pentru finalizarea cercetării. De asemenea, numărul erorilor întâlnite în procesul de introducere a datelor este redus considerabil. Nu în ultimul rând, activitatea operatorului de teren poate fi mai bine controlată. Deşi investiţia iniţială în tablete sau alte instrumente electronice care pot fi utilizate pentru aplicarea chestionarelor este costisitoare, pe termen lung, investiţia se amortizează şi îşi relevă utilitatea. Din ce în ce mai frecvent, chestionarele se aplică şi on-line. Respondentul primeşte un link prin care poate accesa chestionarul pe care îl completează singur. Există o mulţime de soluţii pentru această tehnică, cum ar fi şi cea de tip *open source*, LimeSurvey[1]. Mai putem adăuga aplicarea prin telefon sau e-mail şi, poate mai rar, prin poştă. Pentru detalii despre fiecare în parte, puteţi consulta manuale de metodologie a cercetării sociale sau de piaţă cum ar fi *Marketing Research. An Applied Approach* (Malhotra şi Birks, 2007).

Mulţi studenţi, masteranzi, doctoranzi, cercetători sau chiar firme şi institute de cercetare nu îşi permit achiziţionarea unor tablete cu software dedicat acestor acţiuni. De aceea, utilizează, în continuare, chestionarele tipărite pe hârtie care trebuie introduse într-o bază de date. Apoi această bază de date trebuie curăţată. Soluţii la îndemână în aceste situaţii sunt cele oferite de programele din suita Microsoft Office, mai exact, Microsoft Excel şi Microsoft Access. În proiectele la care am lucrat, de cele mai multe ori, am introdus datele într-o bază de date realizată cu ajutorul Microsoft Access. În continuare, voi descrie paşii prin care realizăm o bază de date pentru introducerea chestionarelor folosindu-ne de Microsoft Excel, apoi de Microsoft Access. Apoi vom afla cum aducem în SPSS datele introduse într-unul dintre aceste programe.

2.1. În ce program introducem chestionarele ?

Acest subcapitol se referă la cercetările în care chestionarele sunt tipărite şi aplicate de un operator de interviu prin procedeul faţă în faţă sau în care acestea sunt completate pe hârtie de către respondenţi.

SPSS are propriile soluţii de introducere a datelor. Mai multe detalii despre acestea şi alte programe din domeniu pot fi găsite pe pagina de internet a producătorului programului[2].

1. https : //www.limesurvey.org/en.
2. http : //www-01.ibm.com/software/analytics/spss.

2.1.1. Introducerea datelor în Microsoft Excel

Microsoft Excel este un program indispensabil în activitatea de cercetare, cu ajutorul căruia putem face diferite calcule, tabele sau grafice. Pe lângă acestea, poate fi folosit şi pentru introducerea într-o bază de date a răspunsurilor la întrebările din chestionare. De fapt, vom alege un software sau un altul în funcţie, în principal, de răspunsul la următoarea întrebare : pot fi evitate erorile de introducere a datelor ? Altfel spus, dacă vrem să introducem răspunsurile la variabila gen, adică 1 = bărbat sau 2 = femeie, putem evita introducerea din greşeală a codului 3 ? Când introduceţi unu-două chestionare, aceasta nu este o problemă deosebită pentru că puteţi observa greşeala neintenţionată. Dar dacă introduceţi 300 de chestionare, fiecare având 100 de variabile, de la un moment dat nu mai observaţi la fel de uşor acest gen de greşeală. Dacă variabila dinaintea sau de după gen include printre codurile valide valoarea 3, atunci chestionarul poate fi introdus decalat. Vom vedea în capitolul 4, dedicat procesului de curăţare, că astfel de erori pot fi identificate, dar corectarea lor presupune timp suplimentar de lucru. Aşadar, dacă pot fi puse condiţii care permit introducerea doar a codurilor corecte, atunci software-ul respectiv este adecvat pentru introducerea datelor. Un alt factor care contează în alegerea programului în care introducem datele constă în posibilitatea de a crea un formular de introducere care este plăcut privirii şi care nu îl oboseşte pe operator.

Să presupunem că avem un chestionar cu trei variabile : id (o variabilă care are un cod unic pentru fiecare respondent), v1 (gen, unde 1 = bărbat sau 2 = femeie) şi v2 (tipul de băutură carbogazoasă preferată, unde 1 = apă, 2 = suc cu cofeină sau 3 = suc de fructe). Vom prezenta în continuare o metodă rapidă de a elabora o bază de date în Microsoft Excel şi de a introduce date în aceasta. De exemplu, la întrebări ne-au răspuns şase persoane. Aşadar, trebuie să avem şase valori diferite la id. Pentru simplitate, acestea vor fi 1, 2, 3, 4, 5 şi 6. Respondentul 1 este bărbat, deci primeşte codul 1. Acesta preferă apa, primind codul 1. Respondentul 6 este femeie, deci primeşte codul 2. Aceasta preferă sucul de fructe, primind codul 3. Datele care trebuie introduse sunt :

id	v1	v2
1	1	1
2	1	2
3	1	2
4	2	1
5	2	3
6	2	3

Deschidem o foaie goală în programul Microsoft Excel, iar cursorul ne duce în celula A1, adică la intersecţia coloanei A cu rândul 1. Coloanele reprezintă variabilele (id, v1, v2), iar rândurile reprezintă respondenţii (cele şase persoane). În celula A1 scriem id. În celula B1 scriem v1. În celula C1 scriem v2. Rezultatul ar trebui să arate ca în figura 2.1.

Figura 2.1. Crearea unei baze de date în Excel. Pasul 1 : introducerea denumirilor variabilelor

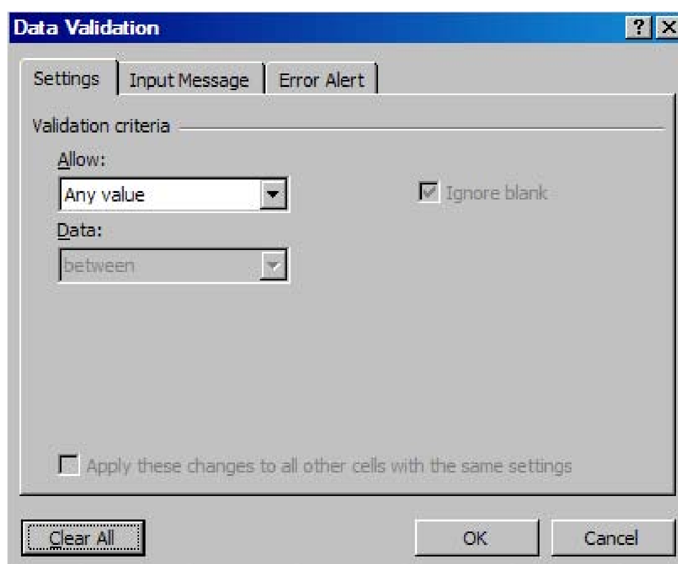
	A	B	C
1	id	v1	v2
2			
3			
4			
5			
6			

Înainte de a introduce datele, trebuie să stabilim condiţiile pentru fiecare variabilă :

- id să aibă valori cuprinse doar între 1 şi 6 ;
- v1 să aibă doar valorile 1 sau 2 ;
- v2 să aibă valori cuprinse doar între 1 şi 3.

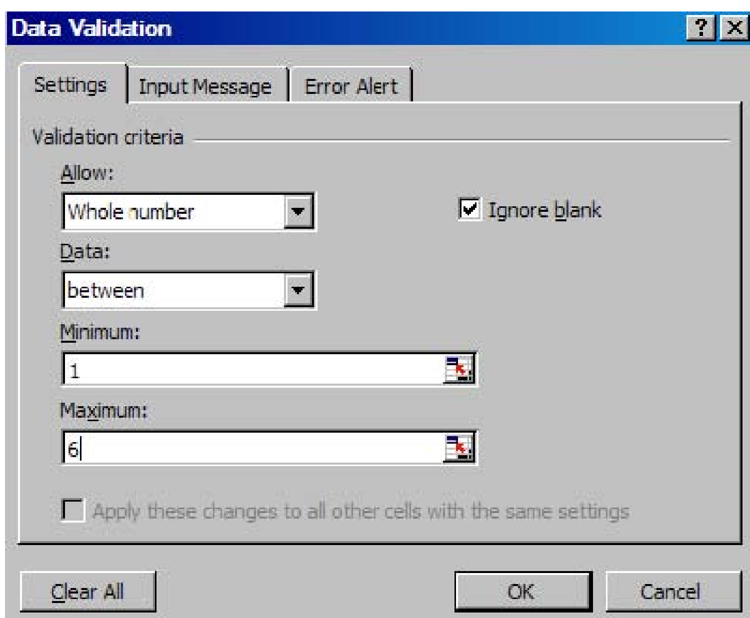
Ne ducem cu mouse-ul deasupra literei A, la numele primei coloane, şi dăm click. Prin această operaţie, selectăm conţinutul întregii coloane A. Apoi mergem în meniul **Data > Validation**. Se va deschide fereastra din figura 2.2. Ne interesează opţiunile din taburile **Settings** şi **Error Alert**. În tabul **Settings** definim condiţia. În tabul **Error Alert** vom scrie un mesaj de atenţionare pentru operatorul care introduce chestionarele, precizând variantele corecte acceptate de celulele respective.

Figura 2.2. Impunerea condiţiilor de introducere a datelor în Excel. Tabul Settings : fereastra iniţială



Pentru că am selectat coloana A, adică variabila id, trebuie să impunem condiţiile pentru aceasta : pot fi introduse doar valorile 1, 2, 3, 4, 5 sau 6. În tabul **Settings**, secţiunea **Allow :** selectăm **Whole number**. Se va activa secţiunea **Data :** în care lăsăm selectat **between**. Pentru că am lăsat selectat **between**, se activează alte două secţiuni, **Minimum :** şi **Maximum :** , în care introducem codul 1, respectiv 6 (figura 2.3). Dacă operatorul introduce din greşeală codul 7, nepermis în acest exemplu, atunci programul îl va avertiza că face o eroare înainte de a-i permite să continue introducerea datelor.

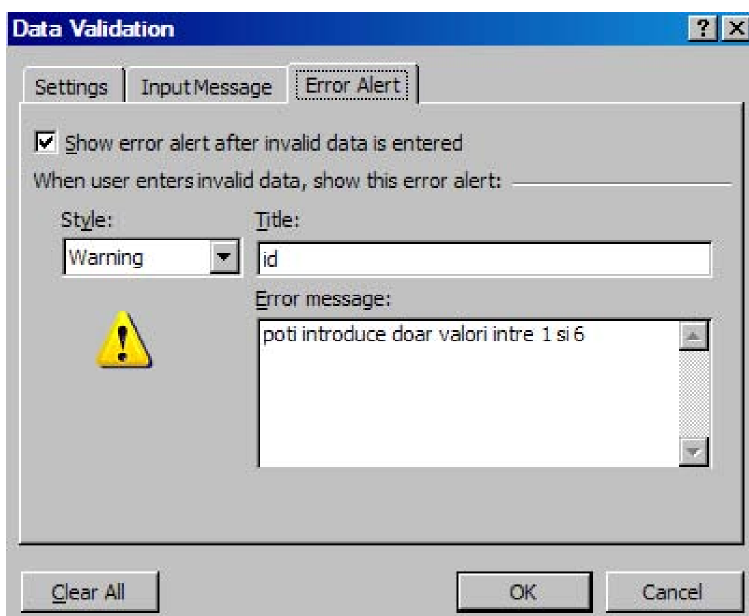
Figura 2.3. Impunerea condiţiilor de introducere a datelor în Excel. Tabul Settings : fereastra cu condiţii



În tabul **Error Alert** :

- în secţiunea **Style**, selectăm **Warning** ;
- în secţiunea **Title** tastăm numele variabilei **Id**. În această secţiune vom introduce numele variabilei pentru care impunem condiţia. Astfel, vom identifica uşor la care variabilă se referă avertizarea ;
- în secţiunea **Error message** tastăm mesajul de avertizare pentru operatorul de introducere : „Poţi introduce doar valori între 1 şi 6” (figura 2.4). Salvăm ceea ce am lucrat (apăsaţi simultan tastele CTRL + S).

Figura 2.4. Impunerea condiţiilor de introducere a datelor în Excel. Tabul Error Alert : fereastra cu mesaj în stilul Warning



Putem trece la variabila următoare, v1. În această etapă, trebuie să instruim programul să accepte doar codurile 1 sau 2. Vom utiliza aceleaşi condiţii ca mai sus. La fel procedăm şi cu v2. Modificăm numele variabilei în secţiunea **Title** şi mesajul care apare în secţiunea **Error message :** Pentru v1, mesajul va fi „Poţi introduce doar valorile 1 sau 2". Pentru v2, mesajul va fi „Poţi introduce doar valorile 1, 2 sau 3". Salvăm ceea ce am lucrat (apăsăm simultan tastele CTRL + S).

Baza de date este finalizată. Acum trebuie să verificăm dacă funcţionează conform aşteptărilor. Verificarea va fi realizată prin introducerea unor coduri greşite. În acest sens, putem opta pentru două variante. Prima variantă, cea mai la îndemână, presupune să tastăm direct în celule, adică introducem valoarea 7 în celula A2. Această variantă este utilă atunci când avem puţine variabile şi puţini respondenţi, adică sunt puţine coloane şi rânduri. Dacă avem foarte multe coloane şi foarte multe rânduri, va deveni obositor pentru operatorul de introducere să le urmărească. Excel oferă o variantă mai simplu de folosit, constând într-un formular. Formularul poate fi accesat din meniul **Data > Form...**. Înainte de a accesa acest meniu, selectăm cele trei coloane care conţin variabilele id, v1 şi v2. În figura 2.5 este prezentat un formular gol şi un formular cu informaţiile introduse pentru respondentul 1. După ce am terminat de introdus datele pentru un respondent, apăsăm butonul **New** şi trecem la următorul respondent.

Figura 2.5. Formular de introducere a datelor în Excel. Formular gol şi formular cu informaţii introduse pentru un respondent

Dacă am fi introdus valoarea 7 la id, atunci programul ne-ar fi avertizat că această valoare nu face parte dintre cele valide. Repetăm operaţiunea şi pentru v1, introducând altceva decât valorile 1 sau 2. La fel şi pentru v3, introducând altceva decât valorile 1, 2 sau 3. Dacă totul funcţionează conform aşteptărilor noastre, atunci putem trece la introducerea datelor. După introducere, baza de date ar trebui să arate ca în figura 2.6. Vom avea şapte rânduri pentru că primul conţine numele variabilelor.

Figura 2.6. Bază de date creată în Excel

	A	B	C
1	id	v1	v2
2	1	1	1
3	2	1	2
4	3	1	2
5	4	2	1
6	5	2	3
7	6	2	3

În practică, lucrurile sunt mai complexe. De exemplu, în foarte multe chestionare, dacă nu în toate, există întrebări-filtru. Adică respondentului 1, pentru că la v2 a ales răspunsul „apă” (codul 1), ar putea să îi fie adresată o altă întrebare : „Ce marcă

preferaţi din lista : 1. Dorna, 2. Bucovina, 3. Izvorul minunilor ? ”. Sau respondentului 2, pentru că la v2 a ales răspunsul „suc cu cofeină” (codul 2), ar putea să îi fie adresată o altă întrebare : „Ce marcă preferaţi din lista : 1. Coca Cola, 2. Pepsi Cola, 3. Adria Cola ? ”. În aceste situaţii, mi se pare mai simplu de utilizat programul Microsoft Access. Preferinţa pentru un program sau altul este, în final, o chestiune de gust sau de experienţă cu unul sau altul. O altă caracteristică după care mă ghidez în alegerea programului cu care lucrez pentru o sarcină anume este uşurinţa cu care pot găsi informaţii ajutătoare despre diferite operaţiuni pe care trebuie să le efectuez cu acesta. Iar Excel, Access şi SPSS stau foarte bine la acest capitol.

2.1.2. Introducerea datelor în Microsoft Access

În cercetările la care am participat, am folosit adesea Microsoft Access, de aceea vă voi explica pe scurt cum se creează o bază de date în acest program.

Baza de date creată în Access sau în Excel va fi importată în SPSS. Pentru ca această tranziţie să funcţioneze corect, dar şi pentru a ne fi uşor să lucrăm cu variabilele din baza de date, trebuie să respectăm câteva condiţii :

- Să citim chestionarul cu atenţie şi să identificăm toate variabilele care trebuie să facă parte din baza de date. O întrebare poate conţine mai multe variabile. Toate variabilele trebuie să se regăsească în baza de date.
- Chestionarele trimise în teren să aibă un identificator (id) unic. De exemplu, id-ul poate fi numărul chestionarului. Dacă avem 1.000 de chestionare de aplicat, atunci acestea sunt numerotate de la 1 la 1.000, fără repetiţii. Numărul chestionarului va fi id-ul. Acesta poate fi şi mai complex de atât, decizia pentru forma finală depinzând de designul cercetării. Cert este că nu există bază de date fără această variabilă.
- Fiecare variabilă să aibă un nume (*name*) care este diferit de al celorlalte variabile.
- Numele să înceapă cu o literă. Numerele pot fi folosite ulterior. Între caractere, fie că sunt litere, fie că sunt numere, nu se lasă spaţiu. Dacă dorim să separăm diferite elemente ale numelui, atunci utilizăm semnul „_”.
  - Corect : *v1*. Incorect : *1v*.
  - Corect : *v1*. Incorect : *v 1*.
  - Corect : *v_1*. Incorect : *v 1*.
- Vă recomand să scrieţi cu literă mică întregul nume. Dacă trebuie să realizaţi o analiză statistică în alt program, iar acel program face distincţia între litere mari şi litere mici, atunci există posibilitatea să vă încurcaţi în denumiri.
- Deşi versiunile mai noi de SPSS permit să folosiţi nume lungi, vă recomand să folosiţi nume scurte, pentru a le putea găsi uşor în lista de variabile din meniuri. Un nume scurt este mai uşor de ţinut minte decât un nume lung.

Dacă va trebui să folosiţi o variabilă cu nume lung într-o analiză efectuată cu alt program de statistică, iar acel program nu acceptă decât, de pildă, maximum 8 caractere, atunci numele va fi trunchiat şi s-ar putea să vă fie greu să o mai găsiţi în baza de date.

Access foloseşte tabele şi formulare create pornind de la tabele. Tabelul este baza de date. Formularul este interfaţa prietenoasă pe care o poate folosi operatorul pentru a introduce chestionarele în baza de date.

Prima etapă în crearea unei baze de date în Access constă în crearea unui tabel. Pentru un chestionar scurt va fi suficient un singur tabel. Pentru chestionare lungi, cu multe întrebări şi, implicit, variabile, va trebui, probabil, să creaţi mai multe tabele. Există multe manuale şi tutoriale dedicate acestui subiect. De aceea voi nota aici doar lucrurile elementare care ne interesează într-o cercetare socială obişnuită.

Să deschidem programul. Odată deschis, mergem în meniul **File > New > Blank Database**. Dăm un nume bazei de date şi o salvăm undeva în computer. Inserăm un tabel în formatul **Design View**. Voi folosi ca exemplu chestionarul utilizat în cercetarea *Diagnoza calităţii vieţii din România* (DCV 2010) realizată în 2010 de Institutul de Cercetare a Calităţii Vieţii din cadrul Academiei Române.

Prima variabilă va fi, întotdeauna, cea care conţine identificatorul unic pentru fiecare chestionar. Chestionarele au fost numerotate de la 1 la *n*, unde *n* reprezintă numărul total de chestionare completate de operatorii de teren conform designului cercetării. Această variabilă poartă numele „nrchest". Aceasta va fi cheia primară (*primary key*) a tabelului. Putem avea o singură cheie primară într-un tabel. Access o va defini singur, dar putem să ne asigurăm că este cea corectă dacă în dreptul variabilei dorite este vizibilă o cheie. Putem alege ca această cheie să fie completată de program sau să o introducem noi. Deşi a doua variantă este mai supusă greşelii, eu o prefer pentru că îmi permite să folosesc chiar informaţia notată pe chestionar. Acest lucru este cu atât mai important atunci când ID-ul nu porneşte de la 1, ci este un cod mai complicat dat de responsabilul de teren fiecărui chestionar. Definirea manuală a cheii primare se face astfel : în tabel, în formatul **Design View**, în coloana **Field Name** introducem nrchest. În coloana **Data Type** selectăm **Number**. Am instruit, astfel, programul că pentru variabila nrchest, introducem numere. Apoi, ducem cursorul pe indicatorul rândului, dăm click dreapta şi selectăm **Primary Key**. Salvăm tabelul (apăsăm simultan tastele CTRL + S).

Acum putem continua definirea variabilelor din chestionar. În chestionarul DCV 2010, respondentului îi sunt adresate mai întâi o serie de întrebări sociodemografice. Echipa care a întocmit chestionarul a avut în vedere, în faza de redactare, faptul că trebuie realizată o corespondenţă perfectă între hârtie şi computer, între chestionar şi baza de date. Astfel, toate variabilele au primit în chestionarul tipărit un nume unic care respectă condiţiile enumerate mai sus.

Prima variabilă are numele d1, a doua d2, a treia d3, iar lista continuă până la d119. După d119, urmează o secţiune scurtă de întrebări adresate operatorului de teren, acestea având numele op1, op2, ..., op9.

Primul lucru pe care respondentul este rugat să îl declare este genul. Numele acestei variabile este d1. Genul (d1) are două variante de răspuns : masculin sau feminin. Varianta masculin a primit codul 1. Varianta feminin a primit codul 2. În baza de date trebuie introduse codurile care se regăsesc în chestionar şi nimic altceva. Transformările se fac, ulterior, în SPSS. De exemplu, dacă în chestionar respondentului i s-a cerut să declare anul în care s-a născut, atunci în baza de date vom introduce anul naşterii. Nu îi vom cere operatorului de introducere să calculeze vârsta şi să introducă valoarea rezultată.

Revenind în Access, în tabelul în format **Design View**, pe următorul rând, sub nrchest, vom introduce în coloana **Field Name** d1, iar în coloana **Data Type** selectăm **Number**. Revin la modul de formatare a chestionarului. Chestionarul este folosit pentru că vrem să calculăm anumite statistici. Statisticile pe care vrem să le calculăm constituie o decizie pe care, teoretic, cercetătorul o ia înainte de a trimite chestionarul în teren. Astfel, vă asiguraţi că se vor culege informaţiile de care aveţi nevoie pentru a răspunde la întrebarea de cercetare. Pentru că statisticile se calculează folosind numere, atunci, în chestionar, când folosiţi aplicarea faţă în faţă cu un operator de teren, din punctul meu de vedere, este obligatoriu să notaţi codurile atribuite variantelor de răspuns :

Corect	**Incorect**
D1. Sexul :	1. Sexul :
1. masculin	☐ masculin
2. feminin	☐ feminin
D4. Statutul ocupaţional :	4. Statutul ocupaţional :
1. salariat	☐ salariat
2. pe cont propriu	☐ pe cont propriu
3. patron	☐ patron
4. zilier	☐ zilier
D26. Cum caracterizaţi calitatea transportului în comun în localitatea dvs :	26. Cum caracterizaţi calitatea transportului în comun în localitatea dvs :
1. foarte proastă	☐ foarte proastă
2. proastă	☐ proastă
3. satisfăcătoare	☐ satisfăcătoare
4. bună	☐ bună
5. foarte bună	☐ foarte bună
Sursa : chestionarul *Diagnoza calităţii vieţii în România 2010*, ICCV.	

Dacă nu notăm codurile, atunci operatorul de introducere a datelor va trebui fie să aloce mult timp înainte de a trece la introducerea efectivă, pentru notarea pe chestionare a codurilor aferente fiecărei variante de răspuns, fie să fie atent ca la fiecare variabilă să introducă corect codul. Se pierde, astfel, timp preţios

şi creşte riscul apariţiei erorilor de introducere. O altă eroare de redactare observată în acest exemplu este atribuirea numelor exclusiv sub formă de număr : în loc de D1, D4 sau D26 (am păstrat numerotarea din chestionar), cercetătorul a atribuit doar 1, 4 sau 26.

Tipul de variabilă (**Data Type**) depinde de caracteristicile informaţiilor conţinute. De regulă, introducem numere şi, uneori, text. Pentru fiecare variabilă definită în **Design View** trebuie să alegem un set de proprietăţi, dintre care le prezint pe cele mai importante :

- **Field Size**. De regulă, vom alege între **Byte**, **Integer** sau **Long Integer**. Diferenţa dintre ele constă în numărul de cifre pe care le poate avea valoarea introdusă.
- **Default value**. Dacă operatorii de teren ar lucra perfect la aplicarea chestionarelor, atunci toate celulele din baza de date ar avea informaţii conform instrucţiunilor chestionarului. Adică ar fi introduse fie răspunsurile valide, fie codurile pentru nonrăspuns. Nonrăspunsul este de trei tipuri : respondentul refuză să răspundă, respondentul nu ştie să răspundă sau întrebarea nu trebuie să îi fie aplicată respondentului. Acestea primesc coduri speciale, diferite semnificativ ca formă de codurile valide. Cele mai utilizate în România sunt 97 = Nu este cazul (NC), 98 = Nu ştiu (NS), 99 = Nu răspund (NR). Există situaţii în care trebuie să le transformăm. De exemplu, o femeie nu vrea să îşi declare vârsta. Operatorul de teren ar trebui să noteze pe chestionar codul 99. Dar 99 poate fi o vârstă validă. Atunci, echipa de cercetare, sub îndrumarea celui care face designul bazei de date, ar putea să instruiască operatorul să noteze pe chestionar codul 999. Aceasta nu mai este o vârstă validă. Dar, dacă ne gândim la salariul lunar, 999 lei poate fi un salariu valid. Atunci, codul de nonrăspuns ar putea deveni –1. Acesta nu mai este un salariu valid. Ideea este să folosim un cod cu totul diferit de variantele de răspuns valide. Punând unul dintre aceste coduri ca **Default Value**, îi spunem programului să introducă singur valoarea respectivă. În acest mod, ne asigurăm că am definit un răspuns uşor de înţeles, când începem analiza statistică. Dacă lăsăm celula goală în tabel, în această fază, s-ar putea să nu mai ştim ce am vrut de fapt să simbolizeze : este o lipsă de răspuns, este o scăpare a operatorului de introducere etc. ? Alegerea codului pentru **Default Value** depinde de tipul întrebării. Dacă răspunsul la întrebare nu depinde de un filtru, atunci vom folosi codul 99 (NR). Dacă răspunsul la întrebare depinde de un filtru, atunci vom folosi codul 97 (NC).
- **Validation Rule**. În acest câmp, introducem o condiţie prin care instruim programul să accepte doar codurile valide înregistrate în chestionar. De exemplu, la d1 avem trei coduri valide : 1 = masculin, 2 = feminin şi 99 = nu răspund. În practică, ultimul cod nu este acceptabil, pentru că operatorul trebuie să vină cu informaţii complete măcar la variabilele sociodemografice esenţiale. Aşadar, regula noastră de validare va fi „1 Or 2 Or 99”. Practic, îi spunem programului să primească doar codurile 1, 2 sau 99. Dacă introducem

codul 3, nu ne va permite să mergem mai departe, deci trebuie să corectăm regula de validare. În chestionarul DCV 2010, la variabila D26, avem cinci variante de răspuns care au primit coduri de la 1 la 5. Am putea scrie „1 Or 2 Or 3 or 4 Or 5 or 99”. Dar, mai simplu, putem scrie : „Between 1 and 5 Or 99”.

- **Validation Text**. Aici putem, opţional, să punem un mesaj ajutător pentru operatorul de introducere. De exemplu : „Variantele corecte sunt 1, 2 sau 99”. Operatorul va identifica mai repede eroarea pe care a realizat-o.

Să recapitulăm luând un exemplu care include şi un filtru în chestionar. După precizarea genului, în DCV 2010, respondentul este rugat să declare care este ocupaţia sa principală actuală. Această variabilă are numele d2, zece variante de răspuns, fiecare având un cod unic şi două tipuri de filtre :

1. agricultor		
2. muncitor (meseriaş)		
3. tehnician, maistru, funcţionar		
4. ocupaţie cu studii superioare		
5. altă ocupaţie	→	Care ? ______________
6. elev, student	→	Dacă 6, sari la întrebarea d6.
7. pensionar	→	Dacă 6, sari la întrebarea d6.
8. casnică	→	Dacă 6, sari la întrebarea d6.
9. acum sunt şomer	→	Dacă 6, sari la întrebarea d6.
10. patron		

Ordinea variabilelor din tabel trebuie să respecte ordinea variabilelor din baza de date. Aşadar, următorul rând în **Design View**, după d1, va deveni d2. La **Default Value** vom introduce 99. La **Validation Rule** vom scrie „Between 1 And 10 Or 99”. La **Validation Text** vom scrie „Poţi introduce doar coduri între 1 şi 10 sau 99”.

Dacă respondentul alege una dintre variantele 1, 2, 3, 4 sau 10, atunci i se va adresa întrebarea următoare : d3. „Din ce an aveţi această ocupaţie ? ”. Dacă respondentul alege varianta 5, atunci va trebui să completeze răspunsul la întrebarea „Care ? ”. Dacă respondentul alege una dintre variantele 6, 7, 8 sau 9, atunci întrebările d3, d4 şi d5 nu i se aplică şi se trece direct la întrebarea d6. Cele trei întrebări nu i se aplică, pentru că se referă la ocupaţie. Aşadar, avem mai multe filtre care trebuie definite şi în baza de date. Folosim filtre în baza de date pentru a grăbi procesul introducerii : introducând valoarea automată 97 la variabilele corespunzătoare, putem sări peste acestea, scutind timp pe care îl putem aloca analizei statistice propriu-zise.

Nu definim filtrele în tabel, ci în formular. Formularul va fi elaborat după ce a fost finalizat tabelul. Adică definim toate variabilele şi proprietăţile lor în tabel, salvăm şi abia apoi trecem la formular. Acum ardem etapele doar în scop didactic.

Să presupunem, aşadar, că am finalizat tabelul introducând toate variabilele din chestionar. Formularul este inserat şi deschis tot în formatul **Design View**. În principiu, toţi aceşti paşi sunt intuitivi în interfaţa programului, motiv pentru care nu mai insist aici. Am creat formularul care conţine toate variabilele din

tabel. Am putea să îi aducem tot felul de îmbunătăţiri estetice, dar aceasta este o chestiune de gust, şi nu de o necesitate imperioasă. Introducerea filtrelor este însă foarte importantă.

Primele două filtre se stabilesc pentru variabila d2 :

- dacă respondentul răspunde cu 5 la d2, atunci programul trebuie să meargă la variabila „Care ? ”.
- dacă respondentul răspunde cu 6, 7, 8 sau 9, atunci programul trebuie să sară peste calupul de întrebări dintre d2 şi d6, mergând direct la d6.

În chestionar, variabila „Care ? ” nu a primit un nume cum ar fi d1, d2, d3 etc. Această situaţie poate fi remediată uşor în program introducând în tabel numele „d2care”. Pentru că respondentului i s-a cerut să precizeze cu propriile lui cuvinte ce ocupaţie are, răspunsurile sunt înregistrate sub formă de text. În câmpul **Data Type** alegem fie **Text**, fie **Memo**. Opţiunea între **Text** şi **Memo** depinde de numărul de caractere care va fi introdus. Pentru simplitate, eu prefer să le definesc pe toate **Memo**. La **Default Value** am introdus 97 pentru că această întrebare se aplică doar celor care au ales codul 5 la d2. Pentru cei care au ales codurile 1-4, respectiv 6-10, această întrebare nu se aplică.

În formular, intrăm în modul **Design View**. Mergem la d2 şi dăm click dreapta pe celulă (nu pe etichetă). Alegem opţiunea **Properties**. Se va deschide o fereastră din care, pentru această situaţie, ne interesează tabul **Event**. Din tabul **Event** ne interesează rândul **On Exit**. Practic, acest eveniment instruieşte programul să aleagă o acţiune în funcţie de codul introdus în d2 atunci când apăsăm tasta **Tab** sau tasta **Enter**, adică trecem la următoarea variabilă din bază. În rândul **On Exit** selectăm **Event Procedure**, apoi dăm click pe cele trei puncte din dreapta celulei. Se deschide o fereastră de cod. Între **Private sub...** şi **End sub**, trebuie să introducem sintaxa :

```
If Me ! [d2] = 1 Then
Me ! [d3].SetFocus
ElseIf Me ! [d2] = 2 Then
Me ! [d3].SetFocus
ElseIf Me ! [d2] = 3 Then
Me ! [d3].SetFocus
ElseIf Me ! [d2] = 4 Then
Me ! [d3].SetFocus
ElseIf Me ! [d2] = 5 Then
Me ! [d2care].SetFocus
ElseIf Me ! [d2] = 6 Then
Me ! [d6].SetFocus
ElseIf Me ! [d2] = 7 Then
Me ! [d6].SetFocus
ElseIf Me ! [d2] = 8 Then
```

```
Me ! [d6].SetFocus
ElseIf Me ! [d2] = 9 Then
Me ! [d6].SetFocus
ElseIf Me ! [d2] = 10 Then
Me ! [d3].SetFocus
ElseIf Me ! [d2] = 99 Then
Me ! [d6].SetFocus
End If
```

Această sintaxă instruieşte programul să respecte filtrele :

- dacă la d2 primeşte codurile 1-4 sau 10, să treacă la variabila d3 pentru că aceasta se aplică acestor respondenţi ;
- dacă la d2 primeşte codul 5, să treacă la variabila d2care, pentru că aceasta se aplică acestor respondenţi ;
- dacă la d2 primeşte codurile 6-9 sau 99, să meargă la variabila d6, pentru că aceasta se aplică acestor respondenţi. Trebuie să definim condiţia şi pentru codul de nonrăspuns.

Salvăm şi închidem fereastra de cod. Ne întoarcem în formular (nu uităm că tabelul este finalizat deja şi nu mai intervenim în el, decât în situaţii excepţionale) şi continuăm cu celelalte variabile, dacă este cazul.

În formular, în modul **Design View**, putem modifica şi estetica formularului. Putem introduce etichete pentru calupuri de întrebări, săgeţi ajutătoare pentru operator etc. Mai important mi se pare să avem în vedere că responsabilul cu elaborarea bazei de date poate lucra cu o versiune mai nouă/veche a programului, iar operatorii de introducere cu una mai veche/nouă a acestuia. Acesta trebuie să asigure compatibilitatea între versiuni. De preferat ar fi să se lucreze pe aceeaşi versiune.

Din punct de vedere estetic, mi se pare important ca formularul să aibă variabilele dispuse în aşa fel încât să încapă pe o jumătate de ecran. Prefer ca introducerea să decurgă de sus în jos, adică variabilele să fie una sub alta :

Varianta preferată de mine					**Variantă posibilă**			
↓	d1		d3		è	d1	↙	d2
	d2		d4			d2care	↙	d3
	d2care		…			d4	↙	d5
	…	↗	…			…	…	…

După ce am realizat formularul şi am introdus toate condiţiile, trebuie să verificăm dacă am lucrat corect. Acest lucru se face simplu, după cum am discutat şi la Excel, introducând în celulele formularului valori ce nu se regăsesc printre răspunsurile valide sau care nu sunt coduri de nonrăspuns. Vom observa imediat dacă filtrele funcţionează sau nu.

2.2. Cum ducem datele în SPSS ?

Am încheiat introducerea datelor. Trebuie să trecem la etapa de curăţare a bazei de date, pe care o realizăm în SPSS. Aşadar, trebuie să ducem datele din formatul Excel sau Access în formatul SPSS. Există mai multe posibilităţi în acest sens, dar, pentru că „paza bună trece primejdia rea", prefer ca, mai întâi, să vizualizez datele în Excel, iar din Excel să le duc în SPSS.

Din Access ducem datele în Excel astfel :

- selectăm tabelul pe care dorim să îl exportăm în Excel,
- deschidem meniul **File > Export**,
- în fereastra care se deschide, la secţiunea **Save as type** alegem unul dintre formatele Excel, de exemplu, **Microsoft Excel 97-2002**, dacă lucrăm cu versiunea 2002 a Access,
- denumim tabelul în modul dorit şi apăsăm tasta **Enter** sau butonul **Export**.

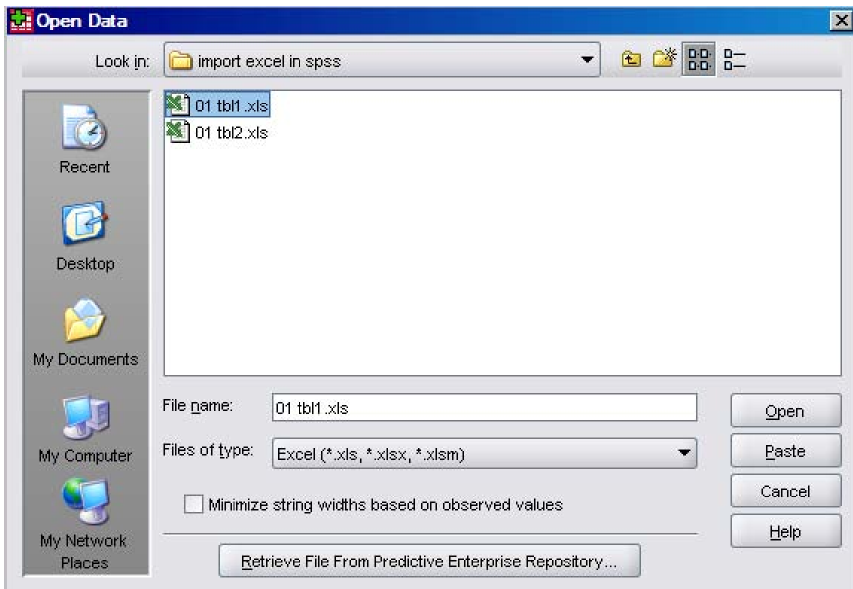
Acum datele sunt în formatul Excel. Pentru a evita erorile generate de modul cum tratează SPSS informaţia venită din alte programe, mai ales în versiunile mai vechi, recomand să verificaţi dacă :

- primul rând din Excel conţine numele variabilelor, iar numele respectă condiţiile enunţate în acest capitol ;
- toate numerele din celule sunt tratate de Excel ca fiind numere. Pentru siguranţă, putem selecta variabilele care conţin numere (coloanele din Excel) dând click dreapta pe numele coloanei şi selectând **Format Cells**. În fereastra care se deschide, în tabul **Number**, la secţiunea **Category** alegem **Number**. De regulă, mai ales în versiunile noi, SPSS citeşte corect informaţia din Excel, aceasta fiind o măsură de precauţie.

Suntem pregătiţi să ducem datele în SPSS. Pentru exemplificare, voi folosi date din DCV 2010. În SPSS pot fi importate baze de date salvate şi în alte formate (de exemplu, fişierele care au extensia **.csv** sunt deseori folosite de analişti) sau chiar în formatul specific altui program de statistică cum ar Stata (fişiere cu extensia **.dta**). SPSS citeşte şi aceste extensii. Lista completă a formatelor recunoscute de SPSS poate fi găsită în documentaţia programului. Mai poate fi utilizat un program comercial, Stat Transfer, care este dedicat acestui gen de operaţiuni. Flexibilitatea este destul de ridicată în domeniul programelor statistice, odată ce ajungi să te familiarizezi cu limbajul acestora.

Deschidem programul şi mergem în meniul **File > Open > Data**. În fereastra care se deschide, selectăm locul unde am salvat tabelele în format Excel. Apoi, în secţiunea **Files of type**, alegem **Excel (*.xls, .xlsx, .xlsm)**. Iniţial, este selectat **SPSS Statistics (*.sav)**. Selectăm tabelul Excel pe care vrem să îl importăm în SPSS (figura 2.7).

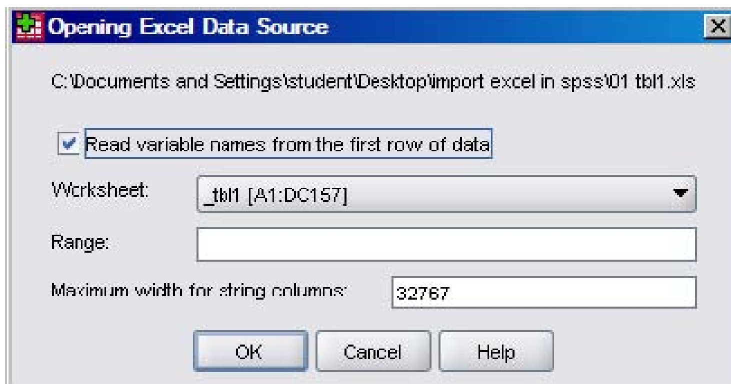
Figura 2.7. Importarea datelor din formatul Excel în formatul SPSS. Selectarea fişierului Excel



După ce apăsăm butonul **Open**, se va deschide fereastra din figura 2.8. Aici trebuie să selectăm foaia de lucru în care sunt datele care ne interesează. De regulă, avem o singură foaie de lucru. Dar dacă avem mai multe şi ne interesează una anume, o vom selecta din listă pe cea corespunzătoare. Înainte de a apăsa OK, verificaţi dacă celula **Read variable names from the first row of data** este selectată. Ar trebui să fie.

Salvăm baza de date rezultată (apăsăm simultan tastele CTRL + S).

Figura 2.8. Importarea datelor din formatul Excel în formatul SPSS. Selectarea foii de lucru care conţine datele



Repetăm aceşti paşi ori de câte ori este nevoie. De exemplu, la DCV 2010 au introdus chestionare mai mulţi operatori. Întrucât chestionarul utilizat are multe variabile, dată fiind complexitatea temei, a fost nevoie de două tabele în Access care cuprindeau, separat, aproximativ jumătate din chestionar. Aşadar, avem de importat în SPSS două baze de date în Excel de la fiecare operator pentru fiecare dintre cei şase operatori. Rezultatul final al procesului de importare trebuie să fie o singură bază de date în SPSS. De aceea, trebuie să parcurgem următoarele etape :

- urmând paşii descrişi anterior, importăm din Excel în SPSS, pentru fiecare operator, cele două baze de date : una cu prima parte a chestionarului şi una cu a doua parte a chestionarului ;
- în cazul fiecărui operator, vom uni în SPSS cele două baze pentru a avea o singură bază, adică întregul chestionar. Rezultă, astfel, şase baze în SPSS pentru toţi operatorii ;
- în fine, vom uni în SPSS cele şase baze într-una singură. Aceasta este baza pe care vom realiza procesul de curăţare.

Să le luăm pe rând. Pentru operatorul M, avem două baze în SPSS. La fel pentru operatorul A, D, E etc. Mai întâi, vom lucra cu cele două baze ale operatorului M. Aceste baze au aceiaşi respondenţi, dar variabile diferite. Am afirmat mai sus că a fost nevoie să împărţim chestionarul în Access, dat fiind numărul mare al variabilelor din chestionarul complex. Pentru a uni aceste două baze, folosim meniul **Data > Merge Files > Add Variables**. Obligatoriu, ambele baze vor avea o variabilă de identificare care ia valori unice pentru fiecare respondent. Fără ea, unirea nu se poate face corect. Aici, această variabilă este numărul chestionarului care, în ambele baze de date, poartă numele nrchest. Deschidem ambele baze de date în SPSS. Primul lucru pe care îl facem este să ordonăm bazele de date, în aceeaşi direcţie, crescător, după nrchest : **Data > Sort Cases >** trecem nrchest în dreapta folosind săgeata **>** lăsăm bifat **Ascending > OK** (figura 2.9). Mai rapid, putem să deschidem **Data View**, dăm click dreapta pe numele variabilei nrchest şi selectăm **Sort Ascending**. Salvăm ambele baze de date (apăsăm simultan tastele CTRL + S).

Acum putem ţine deschise ambele baze de date sau doar pe cea în care aducem noile variabile. Voi explica cum procedăm pentru prima variantă. Deschidem baza primară în meniul **Data > Merge Files > Add Variables**. Se deschide fereastra din figura 2.10. Selectăm cea de-a doua bază. Dacă această bază nu era deschisă, trebuia să selectăm **An external SPSS Statistics data file** şi să căutăm pe computer unde este salvată. Apăsăm **Continue**. În fereastra care se deschide, bifăm **Match cases on key variables in sorted files**, trecem nrchest în căsuţa **Key Variables** şi apăsăm **OK**. Programul ne va avertiza că

trebuia să ordonăm ambele baze de date după variabila-cheie, aici nrchest. Din acest moment, datele din cele două baze se află într-una singură. Dacă dorim să păstrăm bazele iniţiale şi să avem separat baza unită, atunci va trebui să salvăm rezultatul sub o altă denumire.

Figura 2.9. Ordonarea cazurilor : după o variabilă, de la valorile mici la valorile mari

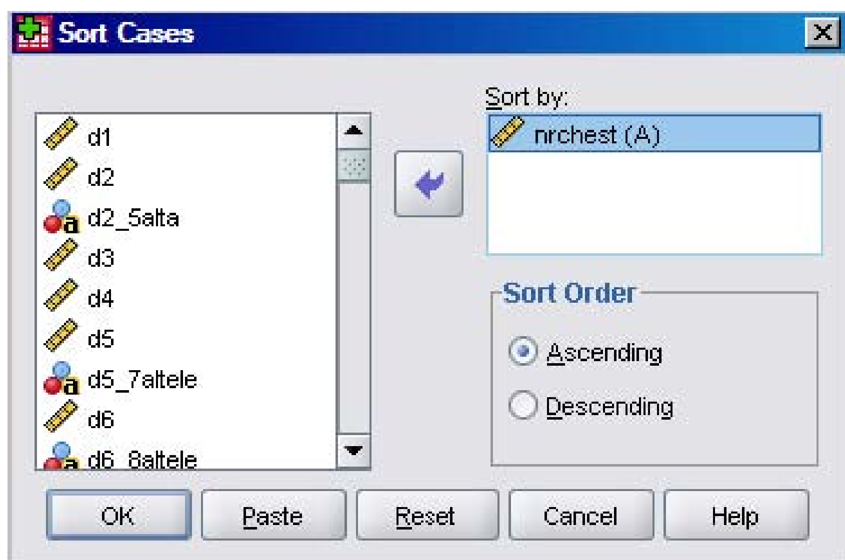
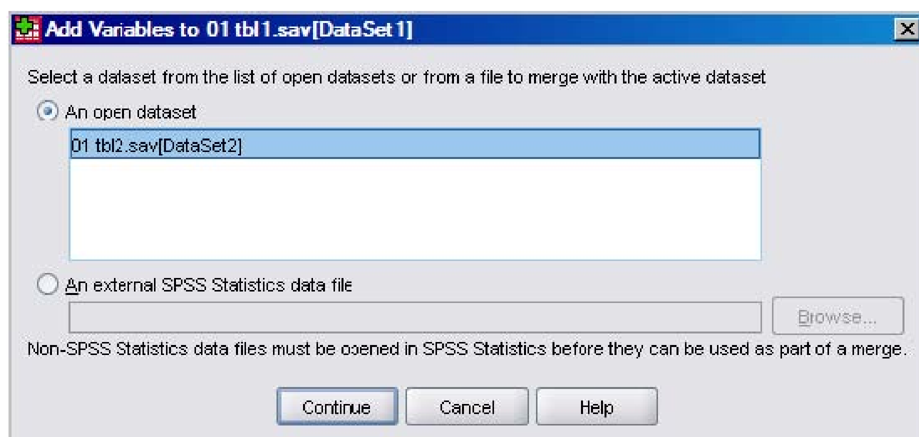
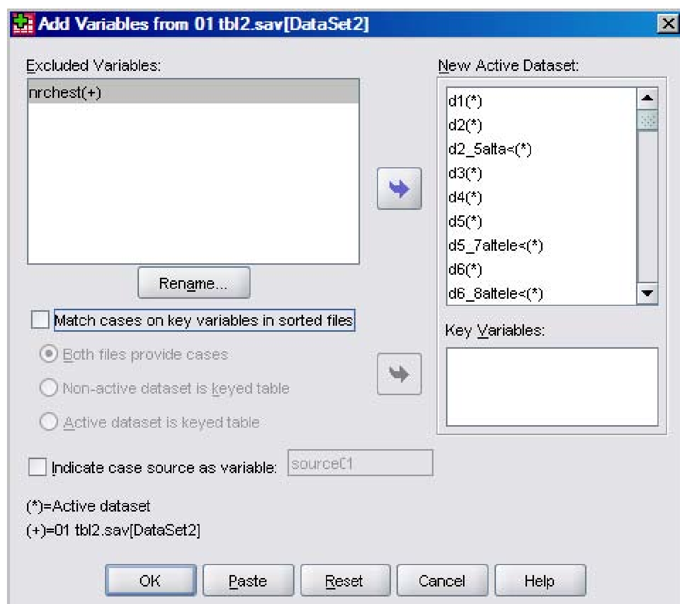
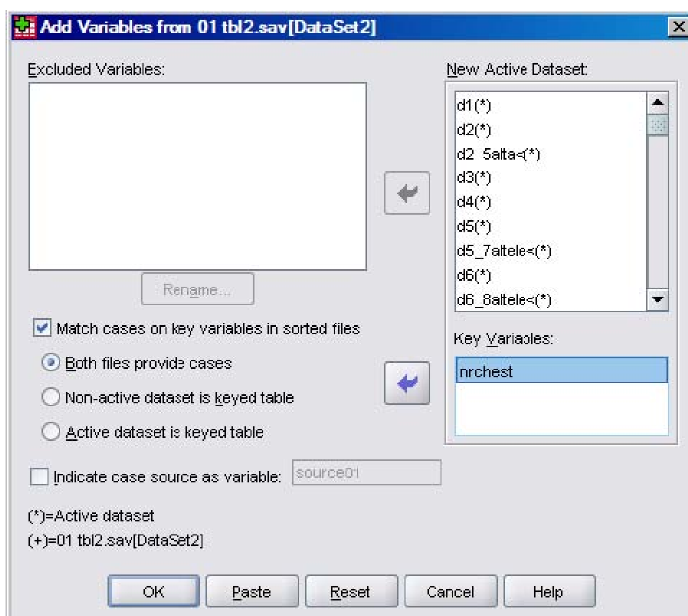


Figura 2.10. Unirea a două baze cu aceiaşi respondenţi şi variabile diferite







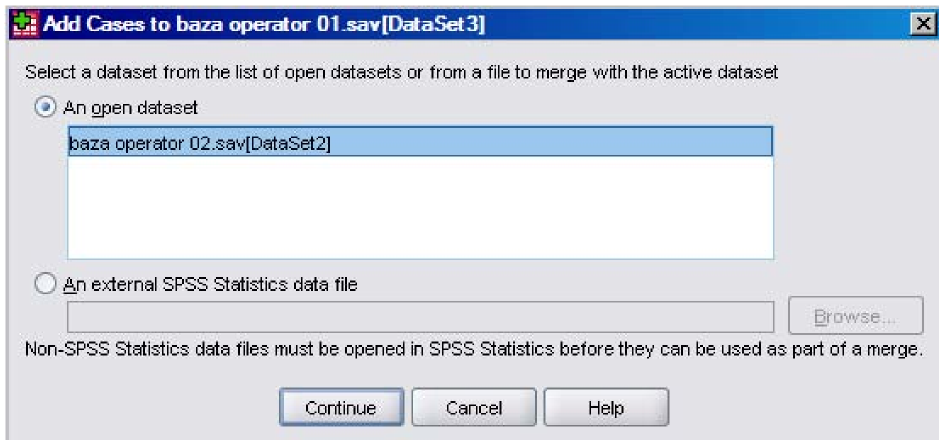
Realizăm această operaţie pentru toţi cei şase operatori de introducere.

După ce am încheiat activitatea, trebuie să unim cele şase baze de date rezultate. De data aceasta, variabilele sunt aceleaşi, însă diferă respondenţii. Vom uni loturi de respondenţi sau, altfel spus, de chestionare. Vom utiliza meniul **Data > Merge Files > Add cases**.

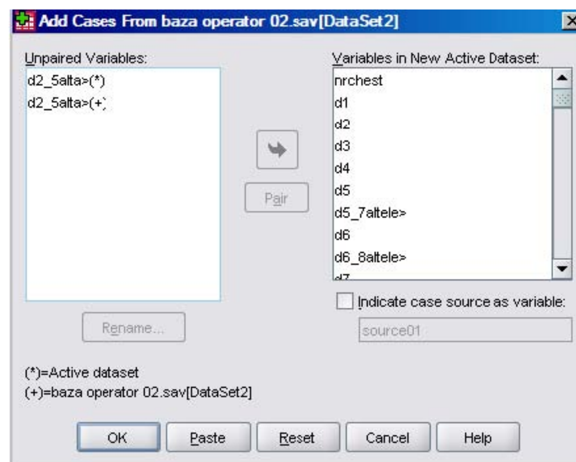
Deschidem baza de date primară, cea în care adăugăm restul de cazuri. În cazul nostru, am denumit operatorii folosind numere, pentru a nu ne încurca în denumiri : baza operator 01.sav, baza operator 02.sav etc. Voi utiliza, ca bază primară, baza primului operator. O deschidem şi o ordonăm ascendent cazurile în funcţie de variabila de identificare, nrchest. Deschidem baza de date a următorului operator. Ordonăm ascendent. Salvăm ambele baze, după această operaţie (apăsăm simultan tastele CTRL + S). Apoi revenim la baza primară, fără să o închidem pe cealaltă. Mergem în meniul **Data > Merge Files > Add cases**. Se deschide fereastra din figura 2.11a. Selectăm baza pe care dorim să o adăugăm în baza primară.

Figura 2.11. Unirea a două baze cu respondenţi diferiţi şi aceleaşi variabile

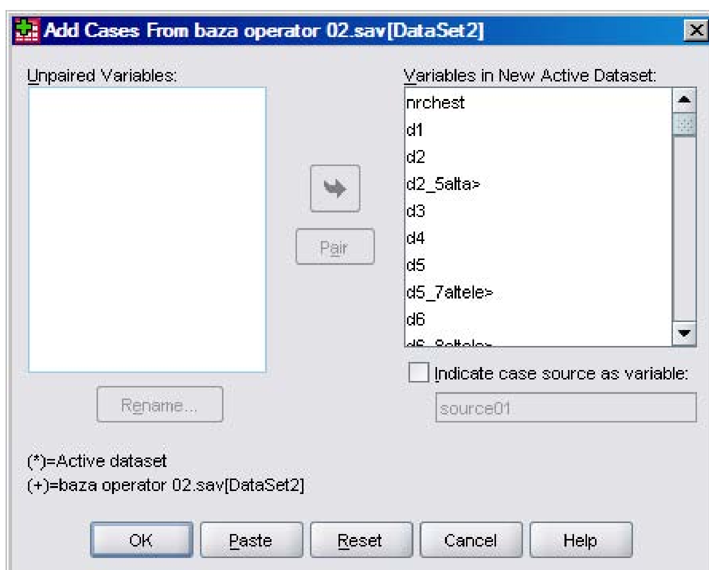
(a)



(b)



(c)



Dacă în secţiunea **Unpaired Variables** (figura 2.11b) apar variabile, înseamnă că nu putem continua unirea. Variabilele care apar în această fereastră au proprietăţi diferite în cele două baze care trebuie unite. Proprietăţile variabilelor pot fi vizualitate în **Variable View**. Renunţăm momentan la unire şi căutăm sursa problemei. În **Variable View**, pentru d2_5alta, în baza operator 01.sav, în coloana **Width** observăm valoarea 765, iar în baza operator 02.sav, observăm valoarea 1000 (figura 2.12). Aceasta este o variabilă care conţine text, mai exact, răspunsul la întrebarea „Care ocupaţie ? " adresată celor care au ales răspunsul „Altă ocupaţie" la d2, „Ocupaţia dvs principală (actuală)". Situaţia se poate repeta pentru toţi operatorii de introducere a datelor. De aceea, înainte de a încerca să unim bazele de date, trebuie să ne uităm în fiecare, în **Variable View** la ce scrie în dreptul variabilelor desperecheate (*unpaired*). Alegem o valoare comună pentru toate cele şase baze, şi anume pe cea mai mare. Operăm modificările şi revenim în meniul de unire. Dacă am lucrat corect, atunci celula **Unpaired Variables** ar trebui să fie goală (figura 2.11c). Apăsăm OK. Repetăm operaţia, până unim toate cele şase baze de la cei şase operatori.

Figura 2.12. Variabile cu proprietăţi diferite (coloana Width din Variable View)

	Name	Type	Width
1	nrchest	Numeric	8
2	d1	Numeric	8
3	d2	Numeric	8
4	d2_5alta	String	765

	Name	Type	Width
1	nrchest	Numeric	8
2	d1	Numeric	8
3	d2	Numeric	8
4	d2_5alta	String	1,000

Din acest moment, putem începe operaţiunea de curăţare şi de validare a bazei de date, acesta fiind subiectul capitolului 4.

2.3. Exerciţii

Pentru aceste exerciţii, utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultate în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a Grupului Românesc pentru Studiul Valorilor Sociale (http : //www.romanianvalues.ro).

1. Deschideţi chestionarul WVS 2012. Alegeţi, la întâmplare, două pagini din chestionar. Răspundeţi la întrebările de pe aceste două pagini.
2. Realizaţi în Excel o bază de date care să corespundă acestor două pagini de chestionar.
3. Introduceţi răspunsurile dvs. în baza de date pe care aţi creat-o.
4. Importaţi baza de date în SPSS.
5. Rugaţi un coleg să vă răspundă la cele două pagini de întrebări selectate anterior.
6. Introduceţi răspunsurile colegului într-o bază de date diferită de cea în care se află răspunsurile dvs.
7. Importaţi baza de date cu răspunsurile colegului în SPSS.
8. Uniţi cele două baze de date.
9. Alegeţi la întâmplare alte două pagini din chestionar. Răspundeţi la întrebările de pe aceste două pagini. Rugaţi acelaşi coleg să vă răspundă şi la aceste întrebări.
10. Realizaţi în Excel o bază de date care să corespundă acestor două pagini din chestionar.
11. Introduceţi în baza de date creată răspunsurile dvs. şi ale colegului dvs.
12. Importaţi baza de date în SPSS.
13. Uniţi această bază de date cu cea obţinută anterior în SPSS.

3. Gestionarea bazei de date

Manipularea şi gestionarea bazei de date presupun un set de cunoştinţe indispensabile analistului. Acesta trebuie să ştie cum se pondereazǎ (*weighting*) o bazǎ de date, cum se filtreazǎ (*select cases*) sau le separǎ (*split file*), cum se agregǎ (*aggregate*) sau se restructureazǎ (*restructure*) etc.

Pentru începǎtori, cred cǎ cele mai importante operaţiuni sunt cele de ponderare, filtrare şi separare. Meniurile aferente acestora şi pe care le discut în acest capitol sunt : **Data > Weight Cases**, **Data > Select Cases** şi **Data > Split File**.

Ponderarea se referǎ la ajustarea bazei de date astfel încât structura eşantionului pentru variabile-cheie sǎ fie similarǎ cu structura populaţiei din care a fost extras acesta şi pentru care dorim sǎ facem inferenţe. Filtrarea este folositǎ atunci când dorim sǎ lucrǎm doar cu anumite cazuri din baza de date sau sǎ realizǎm o nouǎ bazǎ de date, mai restrânsǎ decât cea iniţialǎ. Separarea este folositǎ atunci când dorim sǎ rulǎm o analizǎ pentru grupuri diferite şi sǎ comparǎm rezultatele într-o singurǎ fereastrǎ.

Mai întâi vom prezenta câteva setǎri ale programului care ni se par utile pentru cǎ vǎ ajutǎ sǎ vizualizaţi mai bine informaţia conţinutǎ în baza de date atunci când consultaţi **Outputul**.

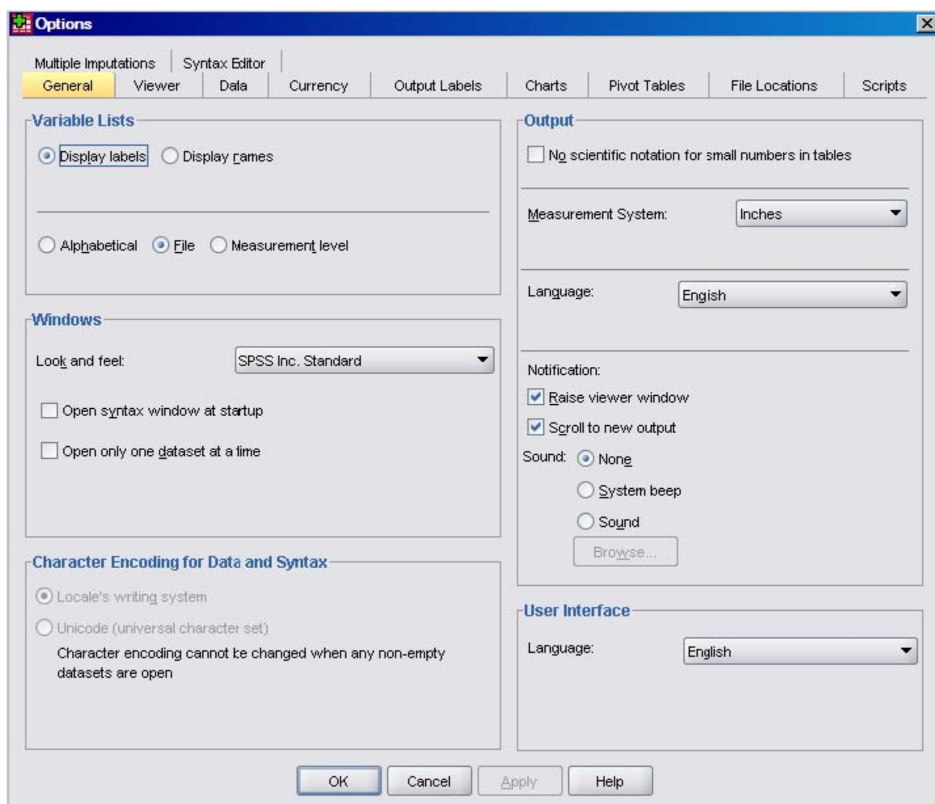
3.1. Câteva setǎri elementare (Edit > Options)

SPSS este apreciat, printre altele, pentru cǎ are o interfaţǎ simplǎ care îi permite utilizatorului sǎ gǎseascǎ rapid lucrurile de care are nevoie. În aceastǎ secţiune, prezint câteva setǎri care cresc uşurinţa cu care se poate utiliza interfaţa. Aceste setǎri pot fi accesate şi modificate în meniul **Edit > Options**. Figura 3.1a prezintǎ fereastra care apare când deschidem acest meniu. Fiind un program complex, şi opţiunile sunt numeroase. Aspectul pozitiv este cǎ ne sunt permise destul de multe intervenţii în opţiunile programului, astfel încât sǎ îl putem ajusta conform nevoilor şi preferinţelor noastre. Cele la care mǎ opresc sunt preferinţele la care am ajuns în timp utilizând programul. Aţi putea avea şi altele pe mǎsurǎ ce dobândiţi experienţǎ cu programul.

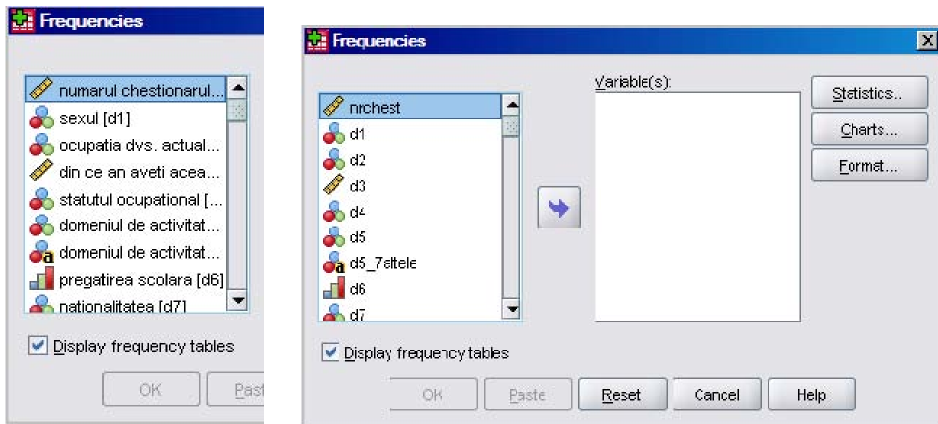
Dintre taburile de opţiuni, ne intereseazǎ urmǎtoarele : **General**, **Output Labels**, **Pivot Tables**, **File Locations** şi **Syntax Editor**.

În tabul **General**, la secţiunea **Variable Lists**, bifăm **Display names**. Iniţial, este bifat **Display labels.** Această operaţiune va permite ca, atunci când deschidem meniurile de analiză, să observăm în lista de variabile numele în locul etichetei (*label*). Putem observa diferenţa în figura 3.1b care prezintă meniul **Analyze > Descriptive Statistics > Frequencies**. În ceea ce mă priveşte, când deschid meniul pentru analize, atunci când văd numele, nu eticheta, îmi este mult mai uşor să găsesc variabilele în lista de variabile. De altfel, putem căuta rapid, după nume, orice variabilă : dăm click în lista de variabile (coloana din stânga) şi tastăm rapid primele două-trei caractere din numele acesteia. În versiunile mai noi de SPSS, putem trece foarte uşor, chiar în interiorul ferestrelor de analize, între nume şi etichete. Mergem în lista de variabile, dăm click dreapta pe oricare variabilă şi alegem **Display Variable Names** sau **Display Variable Labels**, în funcţie de preferinţe şi nevoi (figura 3.1c).

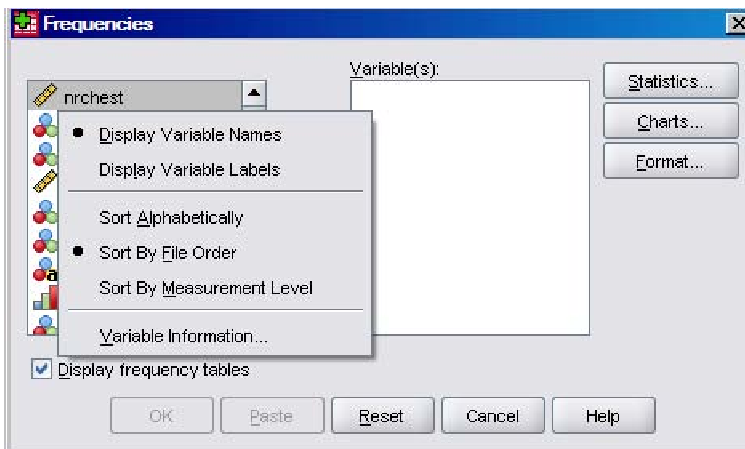
Figura 3.1. Setări care cresc uşurinţa de utilizare a programului. Tabul General (a)



(b)



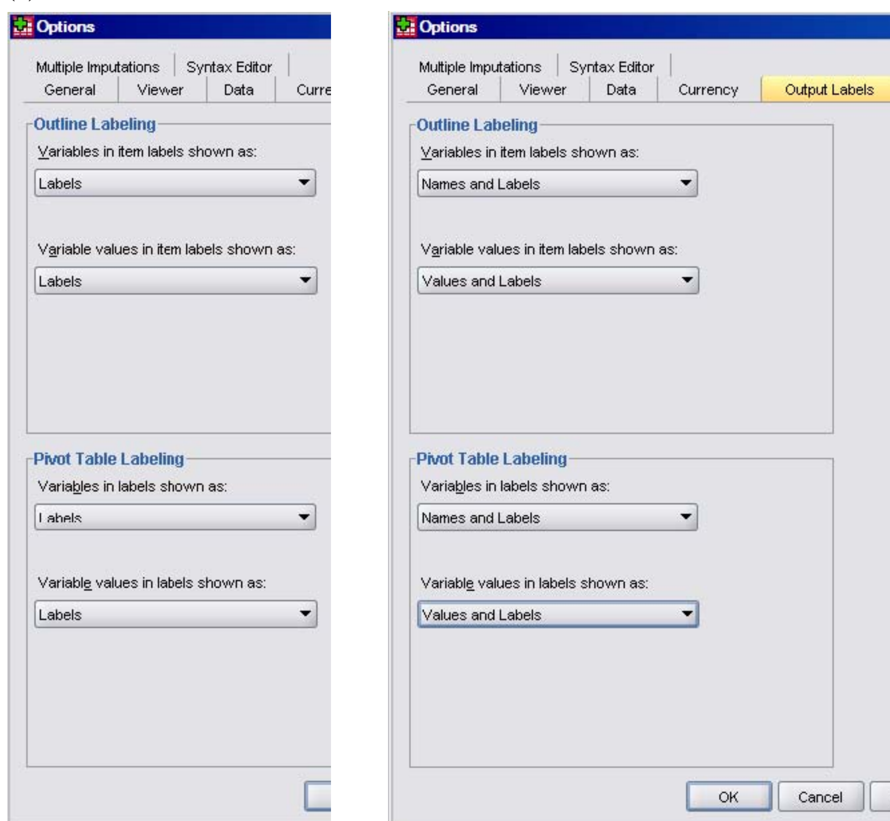
(c)



În tabul **Output Labels** (figura 3.2a) vom selecta, pentru fiecare dintre cele patru câmpuri din secţiunile **Outline Labeling** şi **Pivot Table Labeling**, ambele variante : **Names and Labels**, respectiv **Values and Labels**. Făcând acest lucru, în **Output** vor fi afişate, simultan, atât numele, cât şi eticheta variabilei, respectiv codurile şi etichetele codurilor atribuite variantelor de răspuns. În figura 3.2b este prezentat rezultatul ambelor opţiuni. În al doilea tabel, după ce am modificat opţiunile respective, observăm atât numele, cât şi eticheta variabilei, odată cu codurile şi etichetele atribuite acestora.

Figura 3.2. Tabul Output Labels (Edit > Options) : două tipuri de vizualizare în Output

(a)



(b)

ocupatia dvs. actuala (principala)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	agricultor	259	22.3	22.3	22.3
	muncitori (meserias)	247	21.3	21.3	43.6
	tehnician, maistru, functionar	74	6.4	6.4	50.0
	ocupatii cu studii superioare	106	9.1	9.1	59.1
	elev, student	52	4.5	4.5	63.6
	pensionar	267	23.0	23.0	86.6
	casnica	62	5.3	5.3	91.9
	acum sunt somer	82	7.1	7.1	99.0
	patron	12	1.0	1.0	100.0
	Total	1161	100.0	100.0	

d2 ocupatia dvs. actuala (principala)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 agricultor	259	22.3	22.3	22.3
	2 muncitori (meserias)	247	21.3	21.3	43.6
	3 tehnician, maistru, functionar	74	6.4	6.4	50.0
	4 ocupatii cu studii superioare	106	9.1	9.1	59.1
	6 elev, student	52	4.5	4.5	63.6
	7 pensionar	267	23.0	23.0	86.6
	8 casnica	62	5.3	5.3	91.9
	9 acum sunt somer	82	7.1	7.1	99.0
	10 patron	12	1.0	1.0	100.0
	Total	1161	100.0	100.0	

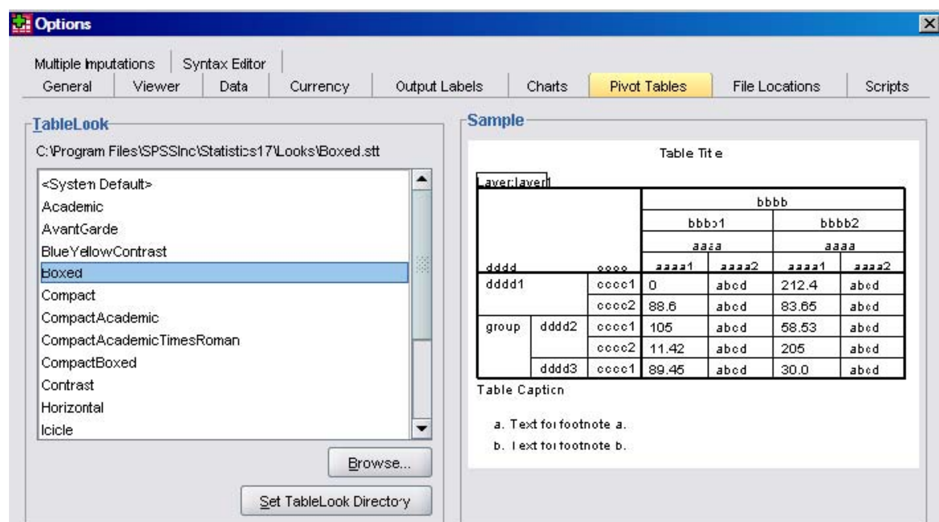
În tabul **Pivot Tables** (figura 3.3a) prefer să selectez, în secţiunea **TableLook**, opţiunea **Boxed**. Tabelul va avea toate celulele delimitate prin borduri. Acest lucru ne ajută, mai ales la tabelele mari, adică cu multe rânduri şi coloane, să citim mai bine informaţia conţinută de acestea. În figura 3.3b se observă diferenţa faţă de modul de prezentare a tabelului din figura 3.2b. Această opţiune este utilă în timpul analizelor. În rapoarte, articole, cărţi sau alte materiale, nu vom copia tabelele din SPSS ca atare, ci le vom realiza în programul de editare a textului pe care îl folosim. Multe tabele oferite de SPSS conţin informaţii ce nu trebuie prezentate ca atare pentru cititor, acestea fiind utile în special analistului. De aceea, aceste informaţii trebuie eliminate sau prezentate în altă formă în cadrul materialului. Putem învăţa să realizăm tabele uşor de citit, dacă parcurgem câteva articole publicate în jurnalele academice din domeniul care ne preocupă. O regulă de bază este : un tabel simplu este uşor de citit. Dacă acesta conţine însă informaţii mai tehnice, atunci punem o notă imediat sub tabel în care explicăm cititorului cum trebuie să citească.

În tabul **File Locations**, la secţiunea **Startup Folders for Open and Save Dialogs**, prefer să bifez **Last folder used** (figura 3.4). O bază de date în format SPSS sau un fişier creat în acest program pot fi deschise fie dând dublu click pe fişier, fie din meniul **File > Open > Data**. O analiză poate dura mai multe zile, în funcţie de complexitatea sa. În a doua zi de lucru, optez pentru a doua variantă de deschidere a fişierului. Dacă bifăm **Last folder used**, atunci, mergând în **File > Open > Data**, programul ne va duce la ultimul fişier utilizat în ultima sesiune de lucru în acest program. Acest lucru este util pentru cei care au multe fişiere pe computer şi, printre acestea, unul dedicat analizelor statistice, fişierul respectiv fiind astfel mult mai rapid de găsit la nevoie. Tot în acest tab, în secţiunea **Session Journal**, ne asigurăm că sunt bifate opţiunile **Record syntax in Journal** şi **Append**.

Programul înregistrează toate operaţiunile noastre. Dacă pierdem sintaxa, ştergând-o din greşeală, o vom găsi în jurnalul pe care îl ţine SPSS. Acest jurnal poate fi salvat în fişierul predefinit de program sau într-un altul, ales de noi.

Figura 3.3. Tabul Pivot Tables (Edit >Options) : modificarea designului tabelelor

(a)

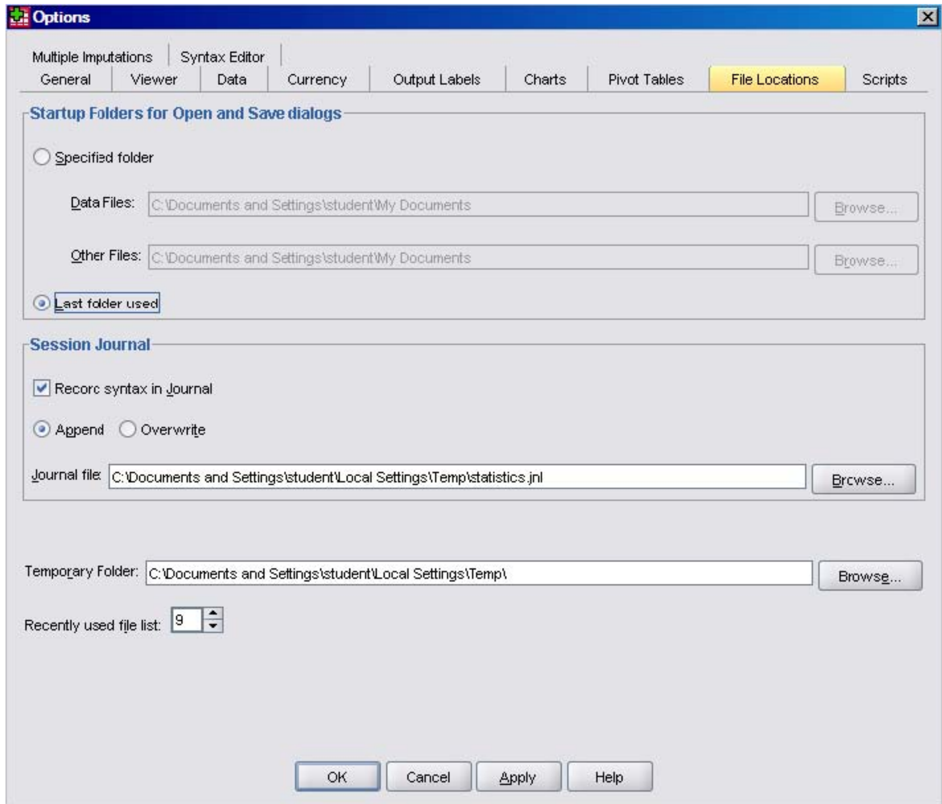


(b)

d2 ocupatia dvs. actuala (principala)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 agricultor	259	22.3	22.3	22.3
	2 muncitori (meserias)	247	21.3	21.3	43.6
	3 tehnician, maistru, functionar	74	6.4	6.4	50.0
	4 ocupatii cu studii superioare	106	9.1	9.1	59.1
	6 elev, student	52	4.5	4.5	63.6
	7 pensionar	267	23.0	23.0	86.6
	8 casnica	62	5.3	5.3	91.9
	9 acum sunt somer	82	7.1	7.1	99.0
	10 patron	12	1.0	1.0	100.0
	Total	1161	100.0	100.0	

Figura 3.4. Tabul File Locations (Edit > Options) : fişierul de lucru şi jurnalul SPSS



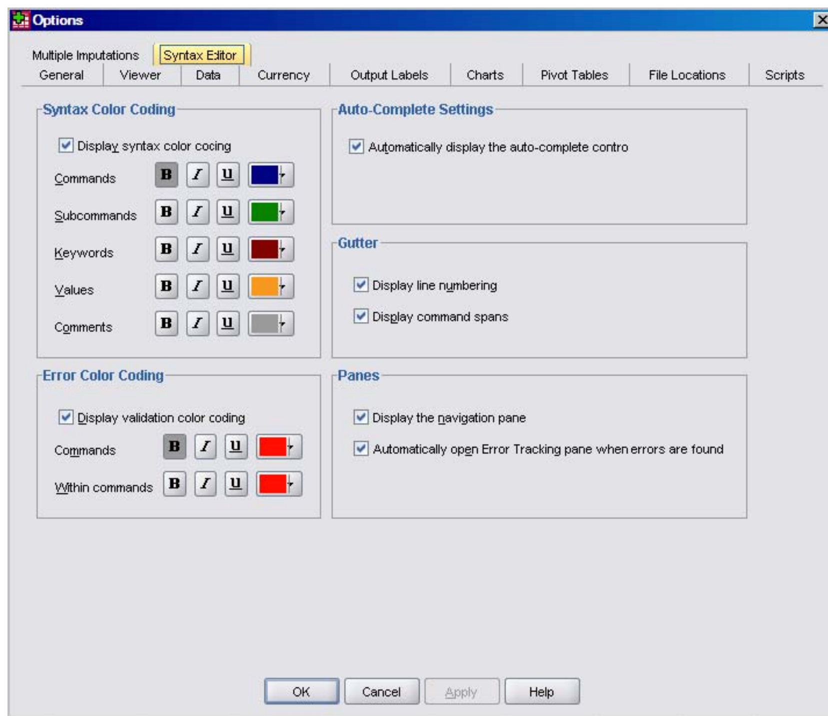
În fine, în tabul **Syntax Editor** (figura 3.5a) putem modifica modul în care arată sintaxa atunci când lucrăm cu ea. În ceea ce mă priveşte, îmi plac culorile alese implicit de program. Mi se pare extrem de utilă opţiunea de completare automată a unei comenzi care poate fi activată bifând **Automatically display the auto-complete control** (figura 3.5b). Dacă dorim să realizăm un tabel de frecvenţă ştiind că începutul comenzii este „fre...”, tastăm „fre...” şi ni se va deschide fereastra din care putem alege comanda corectă. Această opţiune este foarte utilă pentru învăţarea comenzilor uzuale. culorile pe care le afişează editorul sintaxei. Eu modific doar culoarea comentariilor, preferând un gri mai închis. În rest, sunt mulţumit de opţiunile implicite ale programului.

Fereastra sintaxei are două secţiuni : în partea din stânga se află lista comenzilor, iar în partea din dreapta sunt toate comenzile care, rulate, ne vor da analizele dorite. Lista comenzilor ne ajută să navigăm prin sintaxă, când aceasta conţine multe comenzi. Atunci când, din greşeală, am scris o sintaxă greşită, dacă am bifat opţiunea **Automatically open Error Tracking pane when errors are found**, atunci, după cum se observă în figura 3.5c, se deschide o a treia secţiune care ne indică rândul unde se găseşte eroarea, comanda care conţine eroarea şi informaţii

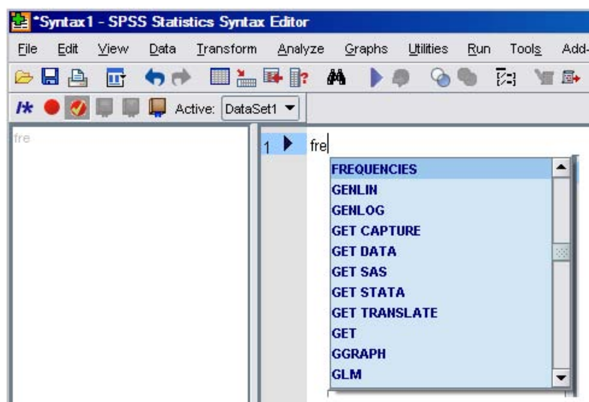
despre eroare. În acest caz, am introdus greşit numele variabilei pentru care dorim să realizăm tabelul de frecvenţă : variabila v1 nu există în baza de date.

Figura 3.5. Tabul Syntax Editor (Edit > Options) : cum putem face sintaxa mai uşor de utilizat

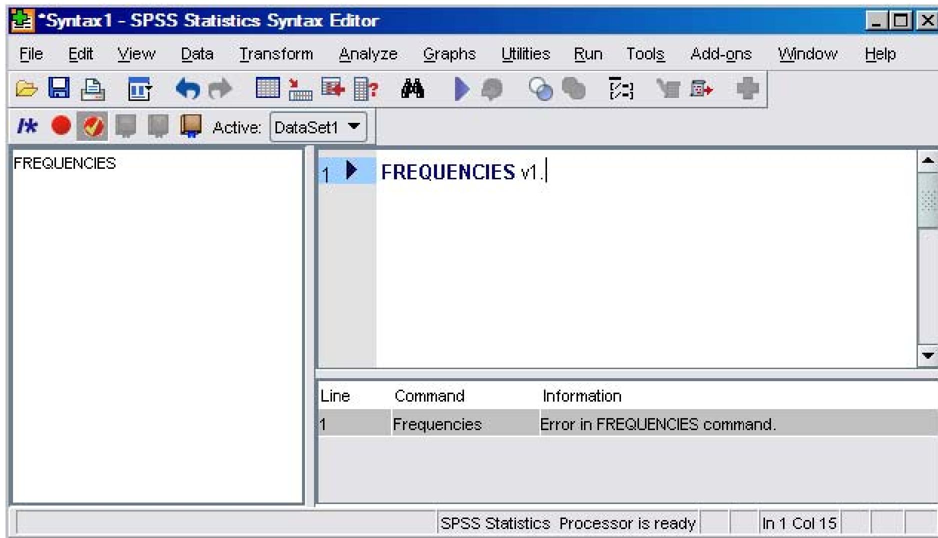
(a)



(b)



(c)



Am prezentat şi opţiunile de la sintaxă pentru utilizatorii începători entuziaşti. Pentru a parcurge această lucrare nu este nevoie să lucrăm cu sintaxa. Pe măsură ce vă familiarizaţi cu programul ar fi util să începeţi să utilizaţi şi sintaxa pe lângă meniuri. Veţi constata că timpul acordat unei analize se reduce considerabil. În plus, veţi avea jurnalul întregii analize la care puteţi reveni oricând pentru referinţe. O parte dintre sintaxele aferente comenzilor utilizate în această lucrare pot fi consultate pe pagina de internet dedicată acesteia.

3.2. Pe scurt, despre structura programului : Data şi Variable View

Aceste două elemente sunt esenţiale în program : **Data View**, respectiv **Variable View**. Probabil că aţi înţeles care este diferenţa dintre ele.

3.2.1. Data View

Data View este secţiunea unde putem vizualiza datele. Dacă baza de date conţine informaţii culese prin aplicarea unui chestionar, atunci fiecare rând va reprezenta un chestionar, iar fiecare coloană va reprezenta o variabilă. O celulă conţine înregistrarea informaţiei pentru un singur individ cu privire la o variabilă. Dacă variabila conţine informaţii despre vârstă, atunci celula pentru rândul 1 va conţine vârsta individului de pe rândul 1. În figura 3.6a este prezentată o secţiune din

baza de date DCV 2010. Rândul 1 reprezintă un român căruia i-au fost puse întrebările din chestionarul DCV 2010. Cifra 1 din dreptul variabilei nrchest pentru acest rând reprezintă numărul alocat acestui chestionar de către cercetător. Observăm că în figura 3.6a este selectată coloana nrchest. Cifra 1 din dreptul rândului 1 şi coloanei d1 reprezintă sexul respondentului la chestionarul cu numărul 1. De unde ştim ce reprezintă d1 ? Dar cifra 1 ? Vom afla proprietăţile acestei variabile în **Variable View**. Să mai zăbovim puţin asupra interfeţei **Data View**.

Figura 3.6. Data View

(a)

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : nrchest 1.0

	nrchest	d1	d2	d3	d4
1	1	1	2	2006	1
2	2	0	1	99	2
3	3	0	1	1955	2
4	4	0	7	97	97
5	5	1	1	99	2

(b)

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : nrchest 1.0

	nrchest	d1	d2	d3	d4	d
1	1	masculin	muncitori (...	2006	salariat	industr
2	2	feminin	agricultor	99	pe cont pro...	agri
3	3	feminin	agricultor	1955	pe cont pro...	agri
4	4	feminin	pensionar	97	97	
5	5	masculin	agricultor	99	pe cont pro...	agri

O parte dintre meniurile SPSS vă sunt familiare pentru că includ comenzi pe care le folosiţi în mod frecvent în alte aplicaţii software uzuale. De exemplu, meniul **File** ne permite să deschidem documente, dar şi să le salvăm. Meniul **Edit** ne permite să copiem (*copy*) şi să lipim (*paste*) diferite elemente. Meniul **Window** ne permite să aranjăm documentele deschise astfel încât să le vizualizăm cât mai pe placul nostru. Meniul **Help** conţine o mulţime de informaţii care ne ajută să înţelegem mai bine programul. Acest meniu se păstrează şi când trecem în **Variable View**. În tabelul 3.1 sunt prezentate unele dintre cele mai utilizate comenzi, care vor fi discutate în această carte.

Tabelul 3.1. Meniuri frecvent utilizate

Meniu	Submeniu	Utilitate
File	New Data New Syntax New Output	Realizăm o bază de date, un fişier de sintaxă sau unul de output, fără informaţii în ele.
	Open Data Open Syntax Open Output	Deschidem o bază de date, un fişier de sintaxă sau unul de output care conţin informaţii.
	Save Save as	Salvăm fişierele pe măsură ce lucrăm. Salvăm fişierele sub alt nume sau în alt loc pe computer.
	Recently Used Data Recently Used Files	Putem deschide un fişier cu care am lucrat într-o sesiune anterioară, fără a-l mai căuta pe computer.
Edit	Insert Variable	Putem introduce manual o variabilă căreia îi definim, ulterior, proprietăţile.
	Go To Case	Putem să găsim rapid un rând din baza de date.
	Go To Variable	Putem să găsim rapid o variabilă din baza de date, dacă îi ştim numele.
	Options	Putem să setăm programul conform preferinţelor personale.
View	Status Bar	Activăm sau dezactivăm **Status Bar**.
	Value Labels	Putem să vizualizăm în **Data View** etichetele atribuite codurilor (figura 3.6b).
	Variables	Trecem din ecranul **Data View** în ecranul **Variable View**.
Data	Identify Duplicate Cases	Putem verifica dacă, după unul sau mai multe criterii, am introdus în baza de date de mai multe ori acelaşi caz. Acest lucru se poate întâmpla, de exemplu, când chestionarele sunt aplicate prin e-mail şi acelaşi respondent ne trimite chestionarul său de pe două adrese de e-mail diferite.
	Sort Cases	Ordonăm cazurile în ordine crescătoare sau descrescătoare în funcţie de una sau mai multe variabile. Putem să ordonăm şi variante combinate.
	Merge Files	Unim două baze de date. Putem uni două baze care conţin aceleaşi cazuri, însă cu variabile diferite, dar şi două baze care conţin cazuri diferite, însă cu aceleaşi variabile.
	Split File	Separăm baza de date după un criteriu. Analiza rulată este prezentată în acelaşi output comparativ pe grupurile definite de criteriul respectiv.
	Select Cases	Activăm sau dezactivăm anumite cazuri astfel încât să rulăm analizele doar pe anumite unităţi. Putem crea baze de date, pornind de la cea iniţială.
	Weight Cases	Ponderăm baza de date. În prealabil, trebuie realizată variabila de ponderare.

Transform	Compute	Realizăm o variabilă nouă, folosind o formulă şi/ sau o funcţie predefinită de SPSS.
	Recode Into Same Variable	Modificăm codurile unei variabile, dar fără a-i modifica structura iniţială.
	Recode Into Different Variables	Modificăm structura unei variabile din baza de date. Rezultatul este o variabilă nouă.
Analyze	Descriptive statistics > Frequencies	Realizăm tabele de frecvenţă, calculăm diferiţi indicatori ai tendinţei centrale, ai variaţiei şi/sau ai poziţionării şi creăm grafice.
	Descriptive statistics > Explore	Explorăm datele. Putem testa asumpţia distribu-ţiei normale folosind indicatori statistici şi grafice.
	Descriptive statistics > Crosstabs	Realizăm tabele de contingenţă, inclusiv testul de semnificaţie *chi square* (hi-pătrat). Calculăm diferiţi indicatori de asociere între variabile categoriale. Putem crea şi graficul specific încrucişării variabilelor categoriale.
	Descriptive statistics > P-P Plots sau Q-Q Plots	Testăm grafic abaterea de la distribuţia normală.
	Compare means > One-Sample T Test	Comparăm media unei variabile din baza de date cu media furnizată de cercetător.
	Compare means > Independent-Sample T Test	Comparăm mediile a două grupuri.
	Compare means > One-Way ANOVA	Comparăm mediile a cel puţin trei grupuri.
	Correlate > Bivariate	Corelăm două variabile metrice.
	Correlate > Partial	Corelăm două variabile metrice, controlând altă variabilă.
	Regression > Linear	Explicăm variaţia unei variabile metrice (depen-dentă), folosind simultan mai mulţi predictori : rulăm analiza de regresie liniară.
	Regression > Curve Estimation	Verificăm dacă între două variabile metrice există o relaţie liniară.
Graphs		Realizăm grafice.
Window	Split	Putem împărţi imaginea în **Data View,** astfel încât să vizualizăm datele în cel puţin două secţiuni. În figura 3.7 este prezentată împărţirea implicită activată prin utilizarea meniului. A nu se confunda cu meniul **Data > Split File**.
Help		Permite accesul la informaţii detaliate despre capabilităţile programului.

Am enumerat în tabelul 3.1 informaţiile pe care un începător trebuie să le acumuleze rapid. După ce acesta le-a înţeles, iar utilizarea lor este deja o rutină, tranziţia către analizele mai complicate devine mult mai uşoară.

Figura 3.7. Meniul Window > Split : rezultatul împărţirii

	nrchest	d1	nrchest	d1	d2
1	1		1	1	2
2	2		2	0	1
3	3		3	0	1
4	4		4	0	7
5	5		5	1	1
6	6		6	0	1
7	7		7	1	1

2	2		2	0	1
3	3		3	0	1
4	4		4	0	7
5	5		5	1	1
6	6		6	0	1

3.2.2. Variable View

În acest meniu creăm variabilele şi le definim proprietăţile. Spre deosebire de **Data View**, în acest caz, rândul este o variabilă, iar coloanele sunt proprietăţi diferite ale acestuia.

Figura 3.8. Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	nrchest	Numeric	8	0	numarul ches:i...	None	None	8	Right	Scale
2	d1	Numeric	8	0	sexul	{0, feminin}...	None	8	Right	Nominal
3	d2	Numeric	8	0	ocupatia dvs. a...	{1, agriculto...	None	8	Right	Nominal
4	d3	Numeric	8	0	din ce an aveti ...	None	None	8	Right	Scale
5	d4	Numeric	8	0	statutul ocupati...	{1, salariat}...	None	8	Right	Nominal
6	d5	Numeric	8	0	domeniul de sc...	{1, agricultu...	None	8	Right	Nominal

Orice variabilă are un nume (coloana **Name**), căruia îi atribuim o etichetă (coloana **Label**). În figura 3.8 observăm, de exemplu, că variabila nrchest are eticheta „numarul chesti…”, iar variabila d1 are eticheta „sexul”, pe când variabila d2 are eticheta „ocupatia dvs. a…”. Putem vedea eticheta întreagă, adică „numarul chestionarului” sau „ocupatia dvs. actuala (principala)” dacă mergem cu mouse-ul între Label şi Values şi tragem de linia care le separă. Observăm că eticheta nu foloseşte diacritice. Uneori, în funcţie şi de setările computerului pe care este deschisă baza de date, acestea nu sunt citite corect, fiind înlocuite cu un simbol cum ar fi semnul de întrebare. De aceea, prefer să am două variante ale bazei de date : una cu diacritice şi una fără diacritice.

Variabilele care au coduri, cum sunt aici sexul (d1), ocupaţia (d2) sau statutul ocupaţional (d4), trebuie etichetate. Acest lucru se face în coloana **Values**. Atunci

când codurile nu sunt etichetate în dreptul variabilei respective, în coloana **Values** apare textul **None**.

În SPSS introducem, de regulă, numere. De aceea, majoritatea variabilelor vor fi numerice (coloana **Type**). Dacă introducem text, atunci tipul se schimbă în **String**.

Pentru că variabilele vizibile în figura 3.8 nu au valori cu zecimale, atunci în coloana **Decimals** ne asigurăm că avem valoarea 0. Dacă o variabilă are valori cu o zecimală, vom înlocui 0 cu 1, iar dacă are valori cu două zecimale, vom înlocui 0 cu 2 ş.a.m.d.

Ar fi indicat ca fiecărei variabile să îi fie definit corect nivelul de măsurare în coloana **Measure**. Astfel, vom beneficia de ajutor suplimentar din partea SPSS care, în anumite meniuri, dacă nivelul de măsurare este definit corect, va sugera diferite modalităţi de lucru.

Aş mai menţiona aici doar coloana **Missing** în care instruim programul, introducând codurile aferente, pentru ignorarea nonrăspunsurilor în analize.

3.3. Ponderarea bazei de date (Data > Weight Cases)

În acest volum pornesc de la asumpţia că datele disponibile sunt culese prin utilizarea unui design de eşantionare probabilist. Un eşantion este probabilist atunci când toate obiectele care fac parte din populaţia de referinţă a studiului au o şansă diferită de zero de a fi selectate în eşantion (Levy şi Lemeshow, 2008). Folosesc cuvântul „obiect", pentru că, în funcţie de nevoile de cercetare, putem fi interesaţi să extragem un eşantion de persoane (români adulţi, cu vârsta egală sau mai mare de 18 ani sau elevi din clasele I-VIII, care fac parte din şcoli în care a fost implementat un program de reducere a abandonului şcolar sau sunt consumatori ai iaurtului cu fructe produs de o anumită companie etc.), dar şi de lucruri (maşini produse de o anumită companie care ies de pe linia de producţie într-o lună, ouă care provin din găini crescute la sol şi ouă care provin din găini crescute în baterii etc.). În toate exemplele fac referire la eşantioane de persoane.

Pentru a extrage un eşantion probabilist, avem nevoie de un cadru de eşantionare. Să presupunem că vrem să extragem un eşantion de persoane adulte cu vârsta de 18 ani şi peste, neinstituţionalizate. Designul frecvent utilizat în România este cel de tip stratificat, multistadial, cu selecţie aleatoare în fiecare stadiu. După ce sunt selectate localităţile, se aleg secţiile de vot şi, în final, cei care vor fi intervievaţi din cadrul fiecărei secţii alese anterior. Informaţiile despre distribuţia populaţiei României, grupată în funcţie de regiunile de dezvoltare, ariile culturale (Sandu, 1999) sau regiunile istorice încrucişate cu mărimea oraşelor şi tipul de sat (aparţinător sau reşedinţă de comună), mărimea satelor după numărul de locuitori sau gradul de dezvoltare al localităţilor rurale pot fi culese de la Institutul Naţional de Statistică[1]. Informaţii

1. Institutul Naţional de Statistică: http://www.insse.ro.

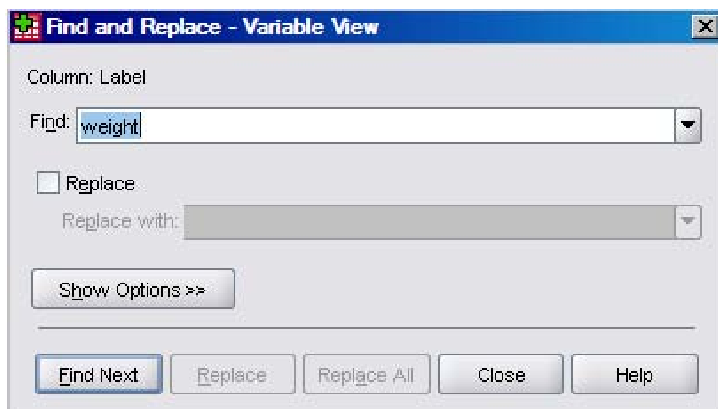
despre secţiile de vot şi membrii acestora pot fi culese de la primăriile localităţilor selectate sau de la Autoritatea Electorală Permanentă[1]. Dacă aceste informaţii sunt actualizate şi armonizate corespunzător, iar operatorii de teren respectă instrucţiunile primite de la cercetători, atunci structura eşantionului pentru variabile-cheie cum ar fi mediul de rezidenţă, sexul, vârsta, educaţia ş.a. ar fi similară cu structura populaţiei. În practică, există situaţii când cele două structuri nu se suprapun perfect. De aceea, se recurge la ponderare. Programul de statistică este instruit să ia în considerare într-o măsură mai mare ceea ce este subreprezentat în eşantion şi într-o măsură mai mică ceea ce este suprareprezentat în eşantion. Acest lucru se face prin construirea unei variabile denumită pondere (*weight*). De exemplu, în cercetarea World Values Survey din 2012 (WVS 2012), al cărui chestionar a fost aplicat şi în România, a fost folosită o variabilă de ponderare. Calcularea ponderilor este un proces destul de laborios care nu face obiectul acestei cărţi. O descriere detaliată, într-un context comparativ, poate fi consultată pe platforma ESS EduNet[2] pusă la dispoziţie în cadrul proiectului European Social Survey[3] sau în lucrările dedicate eşantionării, cum ar fi cea scrisă de Levy şi Lemeshow (2008) care, în *Sampling of Populations : Methods and Applications*, dedică un întreg capitol construirii ponderilor, acesta fiind scris de Paul S. Biemer şi Sharon L. Christ. De asemenea, vă recomand să consultaţi materialele metodologice ale unor cercetări cum ar fi European Values Study[4] sau European Quality of Life Survey[5].

Echipa din România care a aplicat a creat o variabilă de ponderare pentru baza de date. Înainte de a începe analizele propriu-zise, baza de date trebuie ponderată. Există situaţii, ca aceasta la care mă refer, în care baza de date pe care trebuie să o folosim pentru a rula anumite analize statistice nu este creată de noi. Analistul primeşte baza de date pregătită pentru analiză. Aceasta ar trebui să conţină şi variabila de ponderare. În WVS 2012, căutând în **Variable View**, am aflat că variabila de ponderare este V258. Căutarea în **Variable View** se poate face în mai multe moduri. Putem să navigăm, derulând vertical în coloana **Name** şi coloana **Label**, căutând cuvintele-cheie „pondere" sau „weight" sau un alt cuvânt asemănător. Mai rapid ar fi să dăm click în prima celulă din coloana **Label**. Apoi apăsăm iconiţa , cu care suntem familiarizaţi din alte programe utilizate în viaţa de zi cu zi. Sau putem apăsa simultan tastele CTRL + F. Se deschide fereastra din figura 3.9. În secţiunea **Find** tastăm cuvântul „weight". Am ales acest cuvânt-cheie pentru că baza de date WVS 2012 este etichetată în limba engleză, fiind o cercetare comparativă la nivel internaţional, iar datele din

1. Autoritatea Electorală Permanentă : http : //www.roaep.ro.
2. ESS EduNet : http : //essedunet.nsd.uib.no/cms/topics/weight.
3. European Social Survey : http : //www.europeansocialsurvey.org.
4. European Values Study : http : //www.europeanvaluesstudy.eu.
5. European Quality of Life Survey : http : //www.eurofound.europa.eu/surveys/eqls/index.htm.

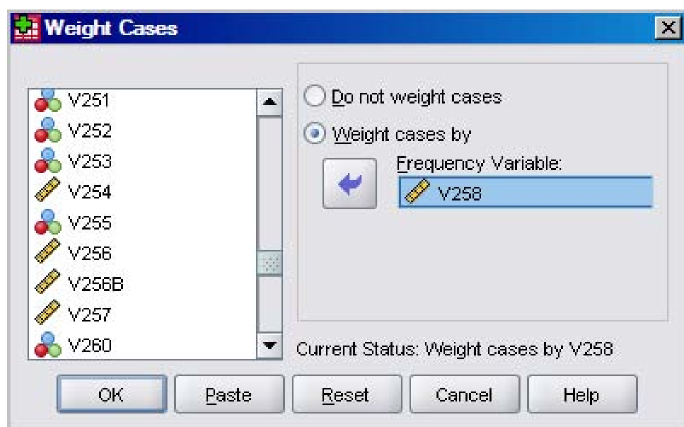
România sunt integrate în acelaşi fişier cu datele din alte ţări. Apăsăm butonul **Find Next** o dată sau de mai multe ori, până când găsim ceea ce căutăm. Dacă etichetele ar fi fost scrise în limba română, am fi folosit cuvântul-cheie „pondere". Nu există o regulă : cuvintele sunt alese în funcţie de ce vrem să găsim. Când nu suntem siguri cu privire la forma sub care este folosit cuvântul, tastăm doar o parte din acesta : „weig" sau „pond".

Figura 3.9. Find : căutare după un cuvânt-cheie



Ponderarea se face din meniul **Data > Weight Cases**. În figura 3.10 este prezentată fereastra cu modificările efectuate, pregătită doar pentru a apăsa butonul **OK**. Căutăm variabila V258 în lista de variabile din stânga. Iniţial, este selectată opţiunea **Do not weight cases**. Dacă baza de date nu trebuie ponderată, această opţiune rămâne neschimbată. Aici bifăm opţiunea **Weight cases by**. S-a activat secţiunea **Frequency Variable** în care introducem, folosind săgeata, variabila V258.

Figura 3.10. Data > Weight Cases : meniul în care activăm ponderea



În output nu se va întâmpla nimic, adică nu va fi produs nici un tabel sau un grafic. În baza de date, indiferent că ne aflăm în meniul **Data View** sau **Variable View**, iar opţiunea **Status Bar** este activată, ar trebui să fie, în colţul din dreapta jos, afişată expresia **Weight On,** ca în figura 3.11.

Figura 3.11. Confirmare vizuală că ponderea este activă



Ponderea poate fi dezactivată în acelaşi meniu, **Data > Weight Cases**, selectând **Do not weight cases** şi apăsând **OK**. Revenind în meniul **Data View** sau **Variable View**, ar trebui să fi dispărut confirmarea **Weight On** prezentată în figura 3.11.

Observăm în figura 3.10 că variabila de ponderare pe care o solicită SPSS este **Frequency Variable**. Acest lucru înseamnă că valorile pe care le ia variabila de ponderare sunt numere de tipul 1, 2, 3, 100, 130 etc. În **Help**, de altfel şi când rulăm diferite analize, suntem avertizaţi că, atunci când cazurile primesc pondere egală cu zero sau ponderi cu numere negative (cu minus), acestea sunt eliminate din analiză. Unele analize acceptă şi ponderi de tipul 1.2, 0.7 etc., iar alte analize nu acceptă deloc ponderi. Trebuie să vă documentaţi bine înainte de a rula o analiză pe o bază ponderată, pentru a vedea în ce măsură este corectată structura eşantionului.

3.4. Filtrarea bazei de date (Data > Select Cases)

A filtra o bază de date înseamnă a selecta din total doar cazurile care îndeplinesc unul sau mai multe criterii. Cazurile care îndeplinesc criteriul de filtrare rămân active în baza de date, iar celelalte sunt dezactivate. De asemenea, putem să le copiem într-o bază de date diferită. Mai putem să ştergem din baza de date iniţială cazurile care nu satisfac criteriul respectiv.

Cercetarea World Values Survey presupune aplicarea unui chestionar cu multe întrebări comune în mai multe ţări într-o perioadă dată de timp. Cercetătorii doresc să compare ţările respective după caracteristicile măsurate în chestionar. După încheierea muncii de teren, vor exista atâtea baze de date câte ţări au fost incluse în cercetare. Aceste baze de date sunt unite într-un singur fişier. De exemplu, în cazul acestei cercetări, puteţi descărca baza de date care conţine toate ţările şi toate etapele din perioada 1981-2005 de pe site-ul World Values Survey, iar, în curând, acesteia îi va fi adăugată şi ultima etapă care, în România, s-a derulat în 2012. Aşadar, avem o bază de date care conţine atât eşantionul românesc, cât şi pe cel german, dar şi altele. Să presupunem că suntem interesaţi să lucrăm

doar cu eşantionul românesc. Pentru că multe cazuri şi multe variabile înseamnă o bază de date mare în termeni de dimensiuni (megabiţi), acest lucru s-ar putea traduce prin durate mai mari de procesare a analizei solicitate computerului pe care lucraţi. Dacă acesta ne permite să lucrăm cu volume mari de date, ne-ar putea totuşi interesa şi partea estetică – să vizualizăm doar variabilele şi cazurile care ne interesează. În oricare dintre aceste contexte, vom utiliza un filtru în baza de date integrată care ne permite să extragem o nouă bază, care să conţină doar eşantionul românesc. Dacă nu vrem să avem mai multe baze de date pe computer, va trebui doar să activăm un filtru care va instrui programul să ia în considerare doar cazurile ce ne interesează, iar după încheierea activităţii care solicita filtrul, îl vom dezactiva şi vom vizualiza, din nou, baza iniţială cu toate cazurile.

3.4.1. Activarea unui filtru : lucrăm pe baza de date iniţială

Să presupunem că vrem doar să păstrăm active anumite cazuri, fără a crea o bază distinctă. Lucrăm doar cu datele culese în România.

Vrem să rulăm o analiză doar pentru bărbaţi : Care este procentul bărbaţilor români care se declară fericiţi sau foarte fericiţi ?

Mai întâi, trebuie să găsim variabila care indică sexul respondenţilor. Căutând în **Variable View**, aflăm că aceasta se numeşte V240. Pentru a activa un filtru în baza de date, trebuie să cunoaştem valorile (codurile) variabilei/variabilelor care constituie filtrul respectiv. Care este codul bărbaţilor ? Pentru a răspunde la această întrebare, realizăm un tabel de frecvenţă (tabelul 3.2) folosind meniul **Analyze > Descriptive Statistics > Frequency**. Pentru a vedea codurile, trebuie să fi făcut modificările în meniul **Edit > Options** aşa cum le-am discutat ceva mai devreme. Codul bărbaţilor (*Male*) este 1.

Tabelul 3.2. Tabel de frecvenţă : Care sunt codurile folosite pentru bărbaţi şi pentru femei ?

V240 Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	723	48.1	48.1	48.1
	2 Female	780	51.9	51.9	100.0
	Total	1503	100.0	100.0	

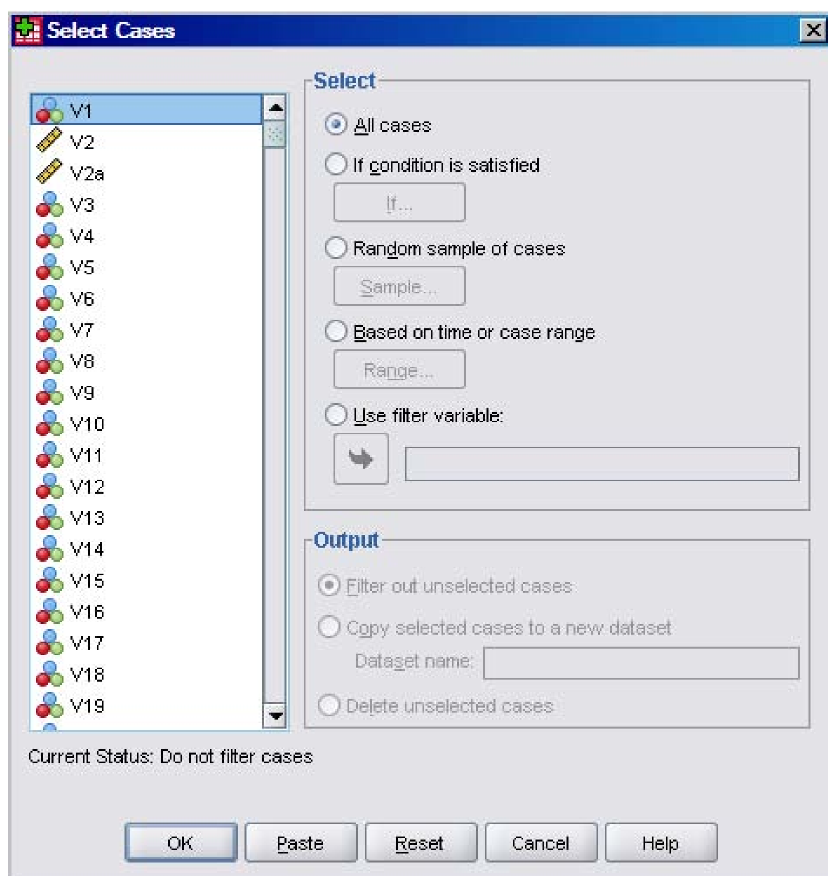
Pentru a selecta doar bărbaţii, trebuie să utilizăm filtrul : **V240 = 1**. Astfel, vor rămâne activi în baza de date doar bărbaţii. Să reţinem structura filtrului : **numele variabilei = cod**. Pentru a activa acest filtru, mergem în meniul **Data > Select Cases** (Figura 3.12a). Iniţial, în secţiunea **Select**, este bifat **All cases**. SPSS utilizează, în această situaţie, toate cazurile din baza de date. Pentru a activa filtrul dorit, trebuie să bifăm **If condition is satisfied**. Observăm că se activează

butonul **If**. După ce am apăsat butonul **If**, se deschide fereastra în care vom pune condiţia prin care instruim SPSS să păstreze activi doar bărbaţii (figura 3.12b). Căutăm variabila V240 în lista de variabile din stânga şi, folosind săgeata, o trecem în secţiunea din dreapta sus. Apoi introducem filtrul : **V240 = 1**. Bărbaţii trebuie să rămână activi. Apăsăm **Continue**. Ne asigurăm că în fereastra iniţială (figura 3.12a), în secţiunea **Output**, este bifată opţiunea **Filter out unselected cases**. Apăsăm **OK**.

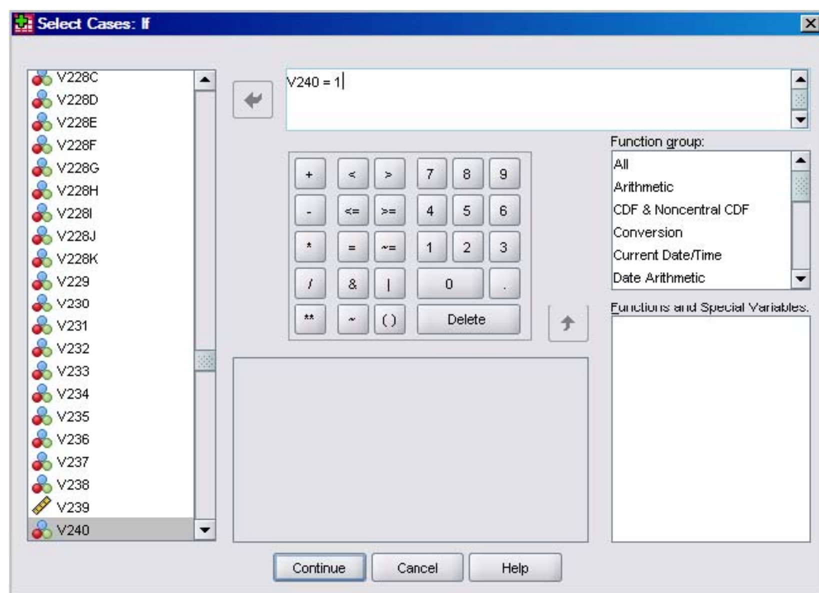
Începătorii cred că ar trebui să utilizeze butoanele pe care le oferă programul (figura 3.12b). Apăsând, de exemplu, semnul „=" şi cifra 1, programul ne va pune automat şi spaţiile necesare între elementele distincte. Semnele „< =" şi „> =" înseamnă „mai mic sau egal", respectiv „mai mare sau egal". Semnul „~=" înseamnă „diferit de". Semnul „&" înseamnă „şi", iar semnul „|" înseamnă „sau".

Figura 3.12. Meniul Data > Select Cases : fereastra iniţială prin care activăm, dezactivăm, copiem sau ştergem cazuri

(a)



(b)



Filtrul este activ. Trebuie să verificăm corectitudinea operaţiunii efectuate. În **Data View** vedem că unele rânduri nu sunt tăiate (2, 3, 6, 7 şi 10), iar altele sunt tăiate (1, 4, 5, 8 şi 9) (tabelul 3.3a). Acesta este modul programului SPSS de a ne spune că un filtru este activat. Dar, pentru a fi siguri că filtrul activ este corect, în această situaţie, realizăm un tabel de frecvenţă pentru variabila folosită în filtru, V240 (tabelul 3.3b). Observăm că doar bărbaţii sunt activi, deci filtrul activ este cel dorit.

Tabelul 3.3. Tabel de frecvenţă : verificarea corectitudinii filtrului

(a)

	V1	V2
~~1~~	6	242
2	6	242
3	6	242
~~4~~	6	242
~~5~~	6	242
6	6	242
7	6	242
~~8~~	6	242
~~9~~	6	242
10	6	242

(b)

V240 Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	723	100.0	100.0	100.0

Nu ne rămâne decât să realizăm un alt tabel de frecvenţă pentru variabila care ne arată procentul bărbaţilor români fericiţi sau foarte fericiţi. Această variabilă poartă numele V10. Folosind meniul **Analyze > Descriptive Statistics > Frequencies**, obţinem tabelul 3.4, unde observăm că 13% sunt „foarte fericiţi” (*Very happy*) şi 58% sunt „destul de fericiţi” (*Rather happy*)[1]. Citim procentele valide (**Valid Percent**) care sunt calculate din totalul bărbaţilor care şi-au declarat nivelul de fericire, adică au răspuns la V10. Acest total este 719 bărbaţi, spre deosebire de totalul general care este 723 de bărbaţi. Folosind procentele cumulate (**Cumulative Percent**), puteam să spunem că 71% dintre bărbaţii români se declarau foarte fericiţi sau destul de fericiţi în 2012.

Tabelul 3.4. Tabel de frecvenţă : Distribuţia fericirii în rândul bărbaţilor români (WVS 2012)

V10 Feeling of happiness					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very happy	91	12.6	12.7	12.7
	2 Rather happy	418	57.7	58.1	70.7
	3 Not very happy	184	25.5	25.6	96.4
	4 Not at all happy	26	3.6	3.6	100.0
	Total	719	99.4	100.0	
Missing	-2 No answer	1	.2		
	-1 Don´t know	3	.4		
	Total	4	.6		
	Total	723	100.0		

În unele situaţii, filtrele de care avem nevoie sunt mai complexe. Folosind aceleaşi date, dorim să aflăm care este nivelul de fericire al bărbaţilor care au educaţie superioară. Filtrul include acum două variabile : sexul şi educaţia. Mai întâi, trebuie să aflăm care sunt variabilele de care avem nevoie pentru analiză. Ştim că sexul este V240. Educaţia este V248. Fericirea este V10. Să ne amintim : am găsit numele variabilelor în **Variable View**, dând click într-o celulă în coloana **Label**, apăsând pe iconiţa reprezentând binoclu şi tastând „sex”, „educ” sau „happ”. Pasul următor presupune să aflăm codurile pe care le vom folosi pentru a crea filtrul. Pentru realizarea acestui obiectiv trebuie să alcătuim un tabel de frecvenţă pentru fiecare dintre cele două variabile de filtrare, sexul (V240) şi educaţia (V248). Deja ştim codurile pentru

1. Traducerea în limba română este preluată din chestionarul românesc al WVS 2012.

sex, aşa că, folosind meniul **Analyze > Descriptive Statistics > Frequencies**, realizăm unul doar pentru educaţie (tabelul 3.5). Dacă, anterior, am creat alte tabele de frecvenţă şi nu am închis baza de date, veţi remarca faptul că în fereastra meniului există acele variabile. Pentru a reveni la setările iniţiale din meniu, apăsăm butonul Reset. Codurile pentru educaţie superioară sunt 8 şi 9. Dacă nu am fi avut etichete pentru coduri, nu am fi ştiut care dintre acestea reprezintă educaţia superioară. Aici este folosită o schemă de clasificare a nivelurilor educaţionale care permite comparaţia între ţări. Puteţi consulta, de exemplu, *International Standard Classification of Education* (*ISCED*)[1] pentru a înţelege mai bine această idee.

Tabelul 3.5. Tabel de frecvenţă: Care sunt codurile pentru bărbaţi şi pentru mediul rural?

V248 Highest educational level attained					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 No formal education	5	.7	.7	.7
	2 Incomplete primary school	13	1.8	1.8	2.5
	3 Complete primary school	26	3.6	3.7	6.2
	4 Incomplete secondary school: technical/ vocational type	27	3.7	3.7	9.9
	5 Complete secondary school: technical/ vocational type	188	26.0	26.3	36.2
	6 Incomplete secondary school: university-preparatory type	120	16.6	16.8	52.9
	7 Complete secondary school: university-preparatory type	165	22.8	23.1	76.0
	8 Some university-level education, without degree	52	7.2	7.3	83.3
	9 University-level education, with degree	119	16.5	16.7	100.0
	Total	715	98.8	100.0	
Missing	–2 No answer	8	1.2		
	Total	723	100.0		

Aşadar, filtrul poate fi scris sub forma: **V240 = 1 & (V248 = 8 | V248 = 9)**. Prima secţiune a filtrului, dinaintea semnului **&**, o cunoaştem: sunt bărbaţii. A doua secţiune se referă la educaţia superioară: observăm că, dacă folosim mai multe condiţii pentru aceeaşi variabilă, trebuie să îi introducem numele de fiecare

1. http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx.

dată. Pentru că un respondent nu poate alege în chestionar la V248 decât un singur răspuns, trebuie să folosim semnul | (sau).

Pentru verificare realizăm un tabel de frecvenţă pentru fiecare dintre cele două variabile de filtrare (tabelul 3.6). Observăm că au rămas active în baza de date doar codurile pentru bărbaţii cu studii superioare.

Tabelul 3.6. Tabele de frecvenţă : verificarea corectitudinii filtrului V240 = 1 & (V248 = 8 | V248 = 9)

V240 Sex					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	171	100.0	100.0	100.0

V248 Highest educational level attained					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8 Some university-level education, without degree	52	30.5	30.5	30.5
	9 University-level education, with degree	119	69.5	69.5	100.0
	Total	171	100.0	100.0	

Un filtru poate fi scris, uneori, în mai multe forme. Încercaţi acest lucru în situaţia dată.

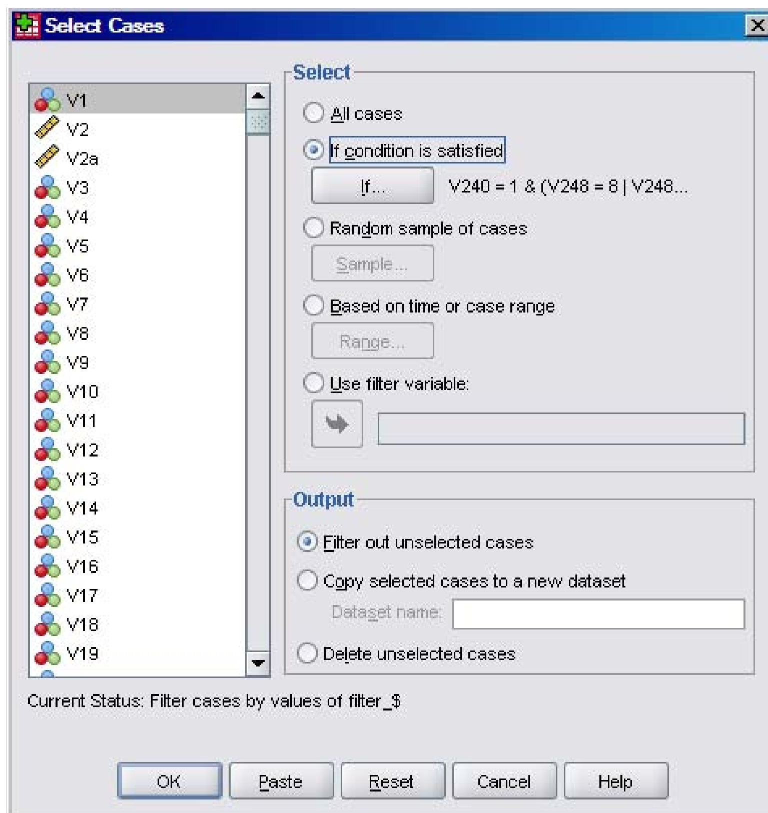
Acum putem rula analiza propriu-zisă : aflarea procentului bărbaţilor români cu studii superioare care se declară fericiţi (tabelul 3.7). 86% dintre aceştia se declară foarte fericiţi sau destul de fericiţi. Procentele sunt calculate din totalul de răspunsuri valide, adică 170. Când filtraţi baza de date, fiţi atenţi la cazurile care rămân active : dacă vă rămân puţine cazuri, atunci trebuie să vă întrebaţi ce relevanţă are analiza respectivă.

Tabelul 3.7. Tabel de frecvenţă : Distribuţia fericirii în rândul bărbaţilor români cu studii superioare (WVS 2012)

V10 Feeling of happiness					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very happy	19	10.8	10.9	10.9
	2 Rather happy	129	75.2	75.7	86.5
	3 Not very happy	22	12.9	13.0	99.5
	4 Not at all happy	1	.5	.5	100.0
	Total	170	99.4	100.0	
Missing	–1 Don´t know	1	.6		
	Total	171	100.0		

Filtrul rămâne activ până când îl dezactivăm. Dezactivarea se face din acelaşi meniu **Data > Select Cases**. Trebuie doar să bifăm **All cases** şi apoi să apăsăm butonul **OK**. Atunci când filtrul este activ, în **Data View** sau **Variable View**, în colţul din dreapta jos observăm pe **Status Bar** Filter On | Weight On. După ce am bifat **All cases** şi am apăsat **OK**, va dispărea **Filter On** din **Staus Bar**. În încheiere, să observăm fereastra meniului cu toate modificările efectuate (figura 3.13). Observăm în dreapta butonului **If** condiţia activă şi, sub lista de variabile şi deasupra butoanelor, expresia **Current Status : Filter cases by values of filter_$**. Această expresie ne indică faptul că SPSS a creat o variabilă care ia valorile 1 şi 0, unde 1 este codul atribuit cazurilor care îndeplinesc condiţia şi 0, codul celor care nu o îndeplinesc. Dacă dorim să reutilizăm filtrul fără a mai face toate aceste operaţiuni, atunci putem redenumi această variabilă în **Variable View** şi, apoi, când avem nevoie de ea, o putem introduce în secţiunea **Use filter variable**. Dacă nu o redenumim, data viitoare când creăm un filtru folosind condiţii noi, aceasta va fi eliminată şi vom pierde informaţia iniţială.

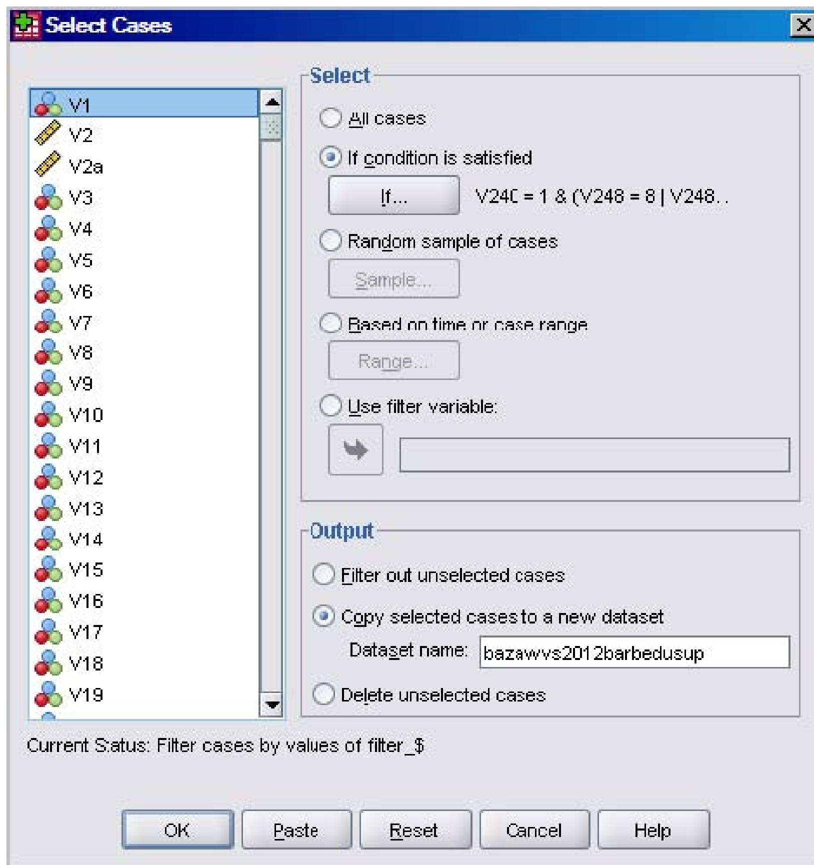
Figura 3.13. Meniul Data > Select Cases : fereastra cu filtrul care menţine active doar anumite cazuri



3.4.2. Crearea unei baze de date folosind un filtru

Folosind acest meniu, putem crea şi o bază de date care conţine doar cazurile ce îndeplinesc anumite condiţii. Singurul lucru pe care trebuie să îl facem este ca, în loc să bifăm **Filter out unselected cases**, să bifăm **Copy selected cases to a new dataset** (figura 3.14).

Figura 3.14. Meniul Data > Select Cases : fereastra cu filtrul care creează o bază nouă de date



Când bifăm **Copy selected cases to a new dataset** se activează opţiunea **Dataset name**. Aici trebuie să introducem un nume pentru noua bază de date, care trebuie să respecte condiţiile impuse numelor variabilelor : să înceapă cu o literă şi să nu conţină spaţii între caractere. Ar fi de preferat să fie şi scurt. Apăsând **OK**, SPSS creează o bază de date care trebuie salvată pe computer, această bază conţinând doar cazurile pe care le defineşte filtrul.

În figura 3.14 se observă că mai avem, în secţiunea **Output**, opţiunea **Delete unselected cases**. Aceasta este utilă doar dacă aţi salvat baza de date originală şi lucraţi pe o copie a acesteia. În caz contrar, veţi pierde informaţii greu de recuperat după această acţiune distructivă.

3.5. Separarea bazei de date (Data > Split File)

Utilizarea filtrelor este un lucru obişnuit în manipularea bazei de date şi în analiza datelor din aceasta.

În unele situaţii dorim să comparăm rezultatul unei analize pentru două sau mai multe grupuri. Care este procentul bărbaţilor foarte fericiţi prin comparaţie cu cel al femeilor fericite ? Predictorii fericirii aleşi în cazul femeilor şi în cel al bărbaţilor funcţionează la fel ? Ideea de bază este că, prin separarea bazei de date (**split file**), putem vizualiza outputul unei analize pentru două sau mai multe grupuri distincte. Pentru aceasta pot fi folosite şi filtre, ceea ce este o chestiune de gust, în multe situaţii.

SPSS ne permite să separăm baza de date în funcţie de o variabilă categorială care conţine cel puţin două grupuri, cum ar fi bărbaţi *versus* femei, locuitori din mediul rural *versus* locuitori din mediul urban, români *versus* germani *versus* bulgari, căsătoriţi *versus* divorţaţi *versus* văduvi etc. Variabila categorială este variabila de separare. Alte variabile vor fi utilizate pentru a rula o analiză pentru fiecare dintre aceste grupuri. În tabelul 3.8 este prezentat tabelul de frecvenţă al variabilei fericire pentru bărbaţi, respectiv, femei.

Lucrăm, aşadar, cu două tipuri de variabile : cea de separare şi cea sau cele pe care le folosim în analize statistice. Aici am separat în funcţie de sex şi am făcut o analiză statistică pentru fericire. Pentru situaţia de faţă, este mai util să realizăm un tabel de contingenţă, despre care vom vorbi în alt capitol al acestui volum. Mi se pare mai utilă această opţiune pe care ne-o oferă SPSS atunci când rulăm un model multivariat, cum ar fi o regresie liniară multiplă. Dacă presupunem că modelul funcţionează diferit pentru bărbaţi şi pentru femei, atunci putem vedea rezultatul în output în funcţie de opţiunea de separare prezentată aici.

Revenind la principiul opţiunii, variabila de separare va fi, întotdeauna, o variabilă categorială : sex, mediu de rezidenţă, stare civilă etc. Putem folosi şi variabile metrice, cum ar fi vârsta, doar dacă aceasta a fost recodificată în prealabil : 18-34, 35-64, 65+. Atunci când recodificăm, creând categorii, trebuie ca numărul de cazuri ce revine fiecărei categorii să fie rezonabil de mare astfel încât să aibă sens comparaţia dintre grupurile rezultate. De aceea, vom folosi variabile de separare cu puţine categorii, mai ales când eşantioanele sunt mici ca volum.

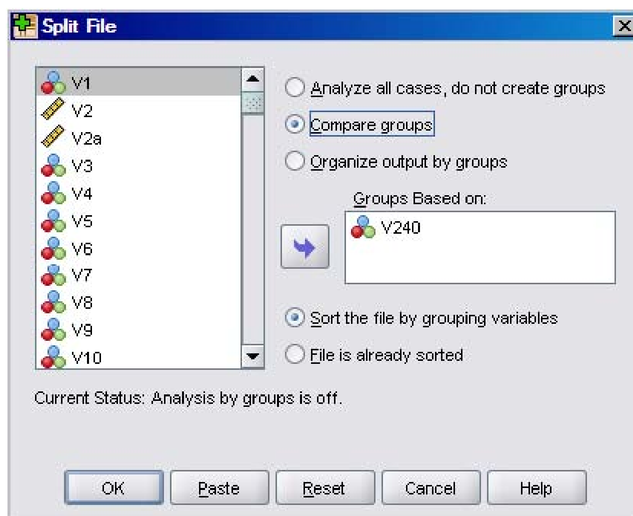
Tabelul 3.8. Reprezentare grafică a rezultatului separării bazei de date

V10 Feeling of happiness						
V240 Sex			Frequency	Percent	Valid Percent	Cumulative Percent
1 Male	Valid	1 Very happy	91	12.6	12.7	12.7
		2 Rather happy	418	57.7	58.1	70.7
		3 Not very happy	184	25.5	25.6	96.4
		4 Not at all happy	26	3.6	3.6	100.0
		Total	719	99.4	100.0	
	Missing	–2 No answer	1	.2		
		–1 Don´t know	3	.4		
		Total	4	.6		
		Total	723	100.0		
2 Female	Valid	1 Very happy	114	14.6	14.7	14.7
		2 Rather happy	415	53.3	53.5	68.2
		3 Not very happy	213	27.3	27.5	95.7
		4 Not at all happy	33	4.3	4.3	100.0
		Total	776	99.5	100.0	
	Missing	–2 No answer	3	.4		
		–1 Don´t know	1	.1		
		Total	4	.5		
		Total	780	100.0		

Dacă dorim să calculăm vârsta medie a bărbaţilor şi a femeilor din România şi să avem această informaţie într-un singur tabel, atunci putem folosi separarea. Variabila de separare va fi sexul, iar variabila pentru care calculăm media va fi vârsta. Variabila sex are numele V240, iar variabila vârstă are numele V242. Mai întâi, alcătuim câte un tabel de frecvenţă pentru a verifica dacă există nonrăspunsuri şi pentru a ne familiariza cu cele două variabile. Variabila sex are două coduri, 1, pentru bărbat şi 2, pentru femeie. Variabila vârstă are foarte multe valori, cea minimă fiind 18 ani şi cea maximă fiind 85 de ani. La variabila sex nu există nonrăspunsuri. La variabila vârstă există trei persoane care nu şi-au declarat vârsta. Observăm că, în baza de date, acestea au fost deja definite ca nonrăspunsuri, pentru că există în tabelul de frecvenţă secţiunea **Missing**, sub **Total**. Celor trei persoane care nu şi-au declarat vârsta le-au fost atribuite codul –2, „nu răspund". Aceste operaţiuni fiind deja realizate, putem trece la analiza propriu-zisă.

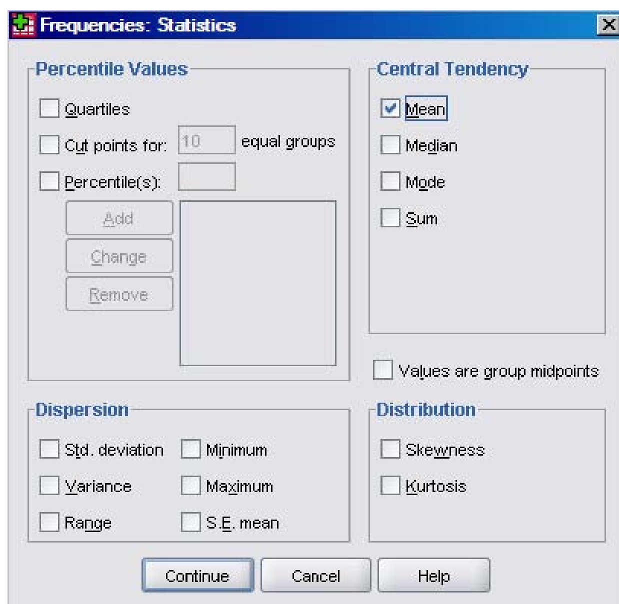
Mai întâi trebuie să separăm baza de date. Acest lucru se face mergând în meniul **Data > Split File** (figura 3.15). În fereastra care se deschide, iniţial este bifată opţiunea **Analyze all cases, do not create groups**. Bifăm opţiunea **Compare groups**. Introducem variabila sex (V240) în secţiunea **Groups Based on**. Apăsăm **OK**. SPSS ne anunţă că este activă opţiunea de separare a bazei de date în Data View sau Variable View, în colţul din dreapta jos pe **Staus Bar** : Weight On | Split by V240 .

Figura 3.15. Meniul Data > Split File



Acum putem calcula media vârstei. Acest lucru îl putem face din meniul cu care ne-am obişnuit deja, **Analyze > Descriptive Statistics > Frequencies**. De data aceasta, vom utiliza şi butonul **Statistics** unde, în secţiunea **Central Tendency**, bifăm **Mean** (figura 3.16).

Figura 3.16. Meniul Analyze > Descriptive Statistics > Frequencies : cum calculăm media unei variabile



Rezultatul analizei este prezentat în tabelul 3.9. Cele două medii sunt 45 de ani pentru bărbaţi, respectiv 47 de ani pentru femei.

Tabelul 3.9. Media vârstei : tabel obţinut prin separarea bazei de date

Statistics			
V242 Age			
1 Male	N	Valid	721
		Missing	2
		Mean	45.00
2 Female	N	Valid	779
		Missing	1
		Mean	47.40

Opţiunea de separare rămâne activă până când o dezactivaţi. Este o situaţie similară cu cea de la filtrare. Trebuie să intraţi înapoi în meniul **Data > Split File** şi să bifaţi opţiunea **Analyze all cases, do not create groups**. Când opţiunea de separare nu mai este activă, textul **Split by** din **Status Bar** dispare.

3.6. Exerciţii

Notă : exerciţiile presupun utilizarea bazei de date European Values Study 2008 România, disponibilă gratuit la ZACAT – GESIS Online Study Catalogue[1].

1. Este baza de date ponderată ? Dacă nu, ponderaţi baza de date.
2. Câţi bărbaţi consideră că prietenii şi cunoştinţele lor sunt importanţi în viaţă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
3. Câte femei consideră că prietenii şi cunoştinţele lor sunt importanţi în viaţă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
4. Câte femei consideră că familia este importantă în viaţă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
5. Câţi bărbaţi consideră că familia este importantă în viaţă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
6. Câţi locuitori ai localităţilor cu peste 100.000 de locuitori consideră că religia este importantă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
7. Câţi locuitori ai localităţilor cu mai puţin de 100.000 de locuitori consideră că religia este importantă ? Aplicaţi un filtru, pentru a răspunde la întrebare.
8. Unde sunt mai mulţi oameni fericiţi : în localităţile cu mai puţin de 100.000 de locuitori sau în localităţile care au peste 100.000 de locuitori ? Separaţi (Split) baza de date, pentru a răspunde la această întrebare.

1. http ://zacat.gesis.org/webview/index.jsp.

9. Cine discută mai frecvent despre politică : bărbaţii sau femeile ? Separaţi baza de date, pentru a răspunde la această întrebare.
10. În ce regiune de dezvoltare sunt cei mai mulţi oameni fericiţi ? Separaţi baza de date, pentru a răspunde la această întrebare.

4. Curăţarea şi validarea unei baze de date

Înainte de a trece la analiza datelor, trebuie să ne asigurăm că acestea nu conţin erori. Aici avem, de fapt, două idei. Una dintre ele este cea pe care o discutăm în acest capitol : eliminarea erorilor de culegere şi de introducere a datelor. Acesta este procesul de curăţare şi de validare a bazei de date. A doua idee se referă la testarea validităţii şi a fidelităţii măsurătorilor cu care lucrăm. În acest sens, putem consulta materiale cum ar fi cele scrise de Mărginean (1982), Saris şi Gallhofer (2007) sau Carmines şi Zeller (1979).

Curăţarea şi validarea unei baze de date constituie un pas esenţial în procesul cercetării cantitative. Acesta este un proces pentru că toate activităţile specifice unei abordări cantitative a socialului sunt interconectate. Cel care primeşte sarcina să cureţe baza de date va comunica permanent cu echipa care a coordonat activitatea de teren. Acesta poate să identifice erori în baza de date care trebuie verificate prin consultarea chestionarului. Curăţarea nu este o activitate făcută într-un birou obscur de cineva care rulează coduri.

În zilele noastre, multe companii de cercetare nu mai tipăresc chestionarele pe hârtie, ci folosesc o metodă de înregistrare digitală. Tableta este un instrument foarte util în acest sens. Folosind această abordare, este redusă considerabil cantitatea de muncă şi de resurse materiale, umane şi temporale necesare pentru finalizarea cercetării.

Informaţiile prezentate aici se aplică atât în situaţiile în care realizaţi o cercetare proprie şi parcurgeţi toate etapele aferente, cât şi în situaţiile în care utilizaţi date culese şi introduse într-o bază de date de altcineva. În a doua situaţie, teoretic, datele sunt deja curăţate, iar baza este validată. În practică, însă, este bine să realizaţi propria verificare : în fond, scăpările altora afectează rezultatul analizelor dumneavoastră.

Etapele esenţiale pentru curăţarea şi validarea bazei de date sunt :

- etichetarea variabilelor şi valorilor variabilelor, acolo unde este necesar acest lucru ;
- dezactivarea nonrăspunsurilor ;
- verificarea introducerii eronate a unor coduri ;
- validarea logică prin urmărirea filtrelor din chestionar, dar şi a unor întrebări factuale ;
- recodificarea unor variabile esenţiale şi construirea unor variabile noi.

Procesul de curăţare ne ajută să apreciem posibilitatea de a utiliza sintaxa. Sintaxa este echivalentul în cod al clickurilor pe care le daţi în meniuri. Din

sintaxă puteţi rula chiar şi comenzi care nu se regăsesc în meniuri. Sintaxa are mai multe avantaje, dintre care aş puncta : (a) avem un jurnal al operaţiunilor pe care le-am realizat în bază, putând reveni oricând la ele pentru a le consulta sau a le rula pe o bază curată ; (b) scade timpul petrecut cu diferite operaţii. Nu trebuie să învăţaţi comenzile. Pe unele dintre ele, pe măsură ce le utilizaţi, le veţi reţine fără probleme. Printre acestea se numără cele pentru tabelul de frecvenţă (**frequencies**), tabelul de contingenţă (**crosstabs**), recodificări (**recode**), pentru realizarea de noi variabile (**compute**) etc. Mai mult, SPSS ne oferă în toate meniurile butonul **Paste** care, apăsat după ce am bifat toate opţiunile dorite, le transformă în coduri pe care le putem salva şi rula oricând.

4.1. Etichetarea variabilelor şi a valorilor variabilelor

Am importat baza de date în SPSS şi s-a deschis fereastra **Variable View** (figura 4.1). Trebuie să completăm informaţii pentru fiecare variabilă (fiecare rând) în coloanele **Label**, **Values** şi **Missing**. SPSS ghidează analistul în anumite situaţii, sugerându-i analizele şi graficele pe care le poate face în funcţie de nivelul de măsurare a variabilelor selectate. Acest lucru este posibil dacă selectăm corect opţiunile din coloana **Measure**. Însă în practică, aceasta este o opţiune pe care o putem ignora, pentru că, pe măsură ce învăţăm să lucrăm cu datele cantitative şi avem mai multe cunoştinţe de statistică, putem decide singuri în situaţiile respective. Este chiar preferabil să controlaţi acţiunile programului, şi nu să îl lăsaţi să ia decizii în locul dumneavoastră.

Figura 4.1. Variable View : baza de date înainte şi după etichetare

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	nrchest	Numeric	8	0		None	None	8	Right	Scale
2	d1	Numeric	8	0		None	None	8	Right	Scale
3	d2	Numeric	8	0		None	None	8	Right	Scale
4	d3	Numeric	8	0		None	None	8	Right	Scale
5	d4	Numeric	8	0		None	None	8	Right	Scale
6	d5	Numeric	8	0		None	None	8	Right	Scale

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	nrchest	Numeric	8	0	numarul chesti...	None	None	8	Right	Scale
2	d1	Numeric	8	0	sexul	{0, feminin}...	None	8	Right	Nominal
3	d2	Numeric	8	0	ocupatia dvs. a...	{1, agriculto...	None	8	Right	Nominal
4	d3	Numeric	8	0	din ce an avet ...	None	None	8	Right	Scale
5	d4	Numeric	8	0	statutul ocupati...	{1, salariat}...	None	8	Right	Nominal
6	d5	Numeric	8	0	domeniul de ac...	{1, agricultu...	None	8	Right	Nominal

În coloana **Label**, introducem explicaţii detaliate despre ce scrie în coloana **Name**. Dacă nu facem acest lucru, atunci când redeschidem baza de date nu vom şti ce înseamnă nrchest, d1, d2 etc. Nu putem găsi nume intuitive pentru toate variabilele. Chiar dacă avem mereu la îndemână un chestionar când lucrăm (de fapt, îl avem), nu este tocmai intuitiv să nu avem etichete în baza de date când

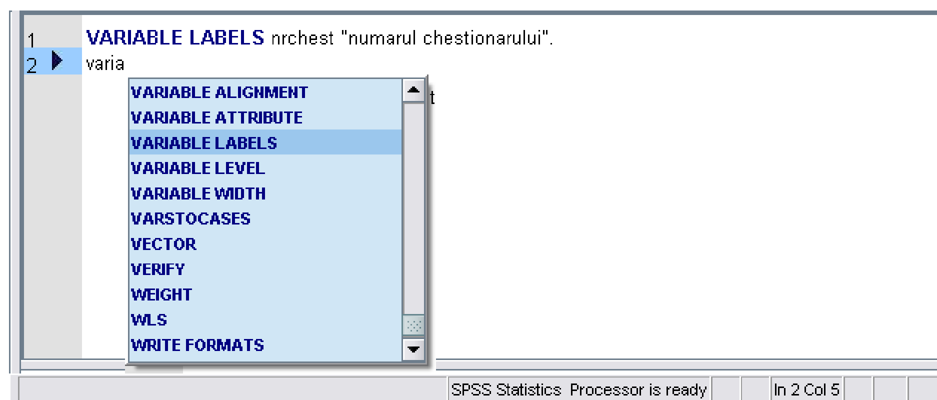
rulăm diferite analize. De regulă, în coloana **Label**, se trece chiar întrebarea din chestionar. Dacă este prea lungă, atunci o putem prescurta alegând cuvintele cele mai importante, astfel încât cei care vor lucra cu această bază de date să înţeleagă uşor informaţiile respective.

Putem proceda în două moduri : (1) în meniul **Variable View** scriem în coloana **Label** în dreptul variabilei care ne interesează sau (2) deschidem un fişier de sintaxă şi scriem comenzile, apoi le rulăm. Recomand varianta a doua pentru că aţi putea şterge din greşeală baza şi nu veţi mai avea sintaxa, caz în care va trebui să o luaţi de la capăt cu etichetarea şi cu celelalte modificări din acest meniu. Un fişier de sintaxă poate fi creat din meniul **File > New > Syntax**. Se va deschide o fereastră nouă similară cu cea din figura 3.5b. Comanda prin care adăugăm o etichetă unui nume de variabilă, adică introducerea unei informaţii în coloana **Label**, este foarte simplă : **VARIABLE LABELS** sau, prescurtat, **VAR LAB**. Iată sintaxa pentru etichetarea celor şase variabile din figura 4.1 :

VARIABLE LABELS nrchest „numarul chestionarului"
VARIABLE LABELS d1 „sexul"
VARIABLE LABELS d2 „ocupatia dvs. actuala (principala)"
VARIABLE LABELS d3 „din ce an aveti aceasta ocupatie ? "
VARIABLE LABELS d4 „statutul ocupational"
VARIABLE LABELS d5 „domeniul de activitate"

Pentru începători, găsesc utilă folosirea denumirii complete a comenzilor. Veţi afla foarte rapid că puteţi prescurta aceste comenzi. Acesta poate fi un exerciţiu de familiarizare cu programul : care este varianta prin care puteţi folosi doar o singură dată comanda **VARIABLE LABELS** pentru toate cele şase variabile ? Folosiţi opţiunea **Help** a programului pentru a afla acest lucru.

Figura 4.2. Fişierul de sintaxă : afişarea listei derulante de comenzi



În figura 4.2 observăm că este suficient să tastăm primele litere din comandă şi programul ne ajută deschizând o listă derulantă din care putem alege ceea ce ne interesează. Nimic mai simplu ! Puteţi întreba : dar de unde ştiu care sunt

comenzile pe care trebuie să le folosesc ? Lăsând acest volum la o parte, puteţi căuta pe internet – SPSS are foarte mulţi utilizatori – şi printre aceştia se găsesc mulţi entuziaşti ai sintaxei. Cu siguranţă, veţi găsi rapid ceea ce doriţi. Programul are un manual bogat care poate fi găsit în meniul **Help** sau chiar pe internet, pe pagina producătorului, IBM.

După ce am scris prima linie de sintaxă, în această situaţie, este suficient să selectăm rândul respectiv, **Copy** şi **Paste** pe rândul următor. Modificăm nrchest cu d1 şi, între ghilimele, scriem eticheta corespunzătoare. Apoi **Paste** pe rândul următor şi modificăm nrchest cu d2 şi, între ghilimele, scriem eticheta corespunzătoare. Repetăm până când am epuizat variabilele care trebuie etichetate.

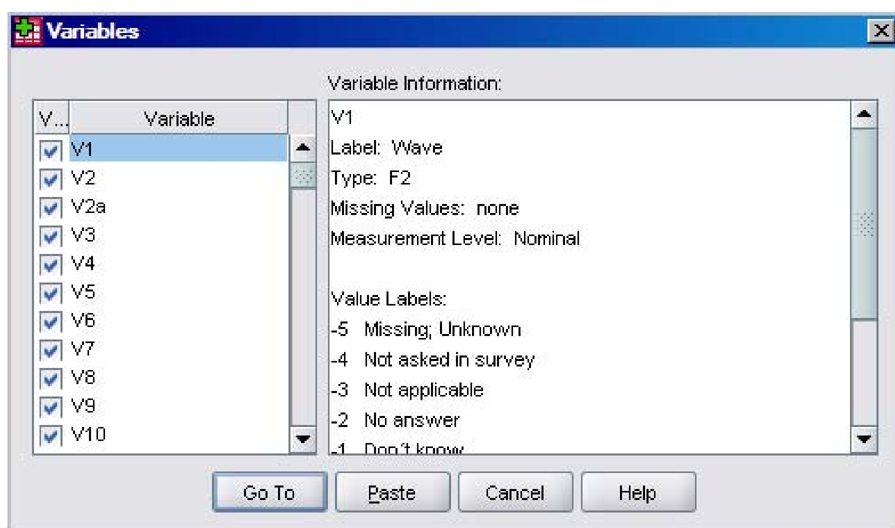
Observăm următoarele :

- putem scrie comanda **VARIABLE LABELS** sau **VAR LAB** fie cu litere mici, fie cu MAJUSCULE. SPSS foloseşte în lista derulantă majuscule, dar acestea nu sunt obligatorii. Pentru a creşte vizibilitatea în interiorul sintaxei, prefer să folosesc pentru comenzi majuscule, iar pentru comentarii litere mici.
- pe rând, între fiecare element al comenzii, lăsăm un spaţiu, apăsând tasta spaţiu. Comandă [spaţiu] numele variabilei [spaţiu] [ghilimele stânga] [eticheta] [ghilimele dreapta] [punct].
- după comanda **VARIABLE LABELS**, notăm numele variabilei, aici nrchest sau d1 sau d2 etc. SPSS oferă posibilitatea de a pune automat numele variabilei în fişierul de sintaxă. Putem merge în meniul **Utilities > Variables** (figura 4.3). Variabila nrchest este prima. Nu ne va fi de mare folos. Dar să presupunem că vrem să găsim rapid variabila V240 : dăm click în lista de variabile din stânga ferestrei pe oricare variabilă, astfel încât aceasta să fie selectată (de exemplu, aici este selectată V1). Apoi tastăm rapid primele două-trei caractere din numele variabilei care ne interesează, aici V240. Programul ne va duce imediat la variabila V240. Ne asigurăm că este selectată şi apăsăm butonul **Go To**, dacă vrem să fie afişată în baza de date, sau butonul **Paste**, dacă vrem să fie copiată în sintaxă. Vom alege a doua opţiune. Acest meniu este foarte util atunci când variabilele nu au denumiri atât de intuitive ca d1, d2, V240 etc., ci mai greu de ţinut minte, cum ar fi tvtot, trstlgl, prtvtcbe etc., acestea fiind denumiri folosite în baza de date a cercetării *European Social Survey 2012*[1].
- după numele variabilei, între ghilimele, scriem eticheta. Ghilimelele, în principiu, sunt necesare dacă eticheta conţine caractere speciale cum ar fi cratima, semnul exclamării, punct etc. În plus, delimitează vizual sintaxa.
- întreaga comanda se încheie cu punct.

---

1. http : //www.europeansocialsurvey.org/docs/round6/survey/ESS6_appendix_a8_e01_0.pdf.

Figura 4.3. Meniul Utilities > Variables : cum găsim rapid o variabilă şi cum îi copiem numele în sintaxă

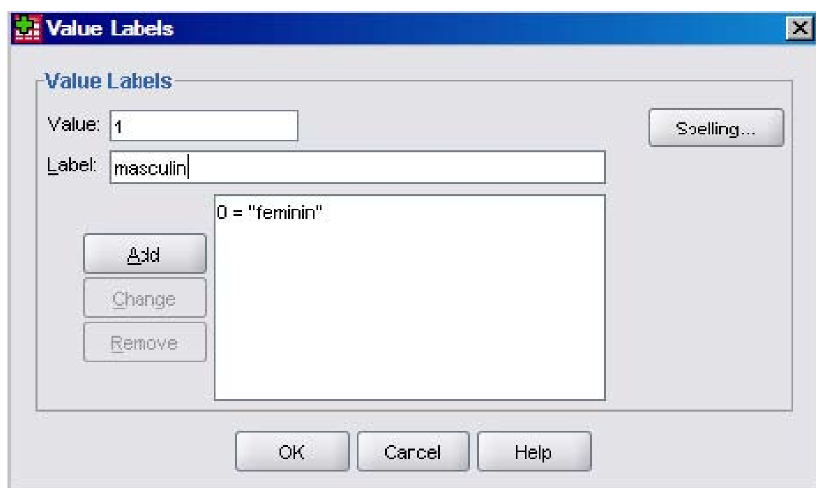


După ce am încheiat de notat sintaxa şi am salvat fişierul, putem rula sintaxa. Deschidem meniul **Run**, unde există mai multe posibilităţi. Dacă vrem să rulăm doar o anumită comandă, şi nu întregul fişier de sintaxă, atunci alegem **Selection**. Pentru a vedea modificările, mergem în **Variable View** (figura 4.1). Puteţi rula sintaxa şi fără să utilizaţi acest meniu : găsiţi prescurtarea !

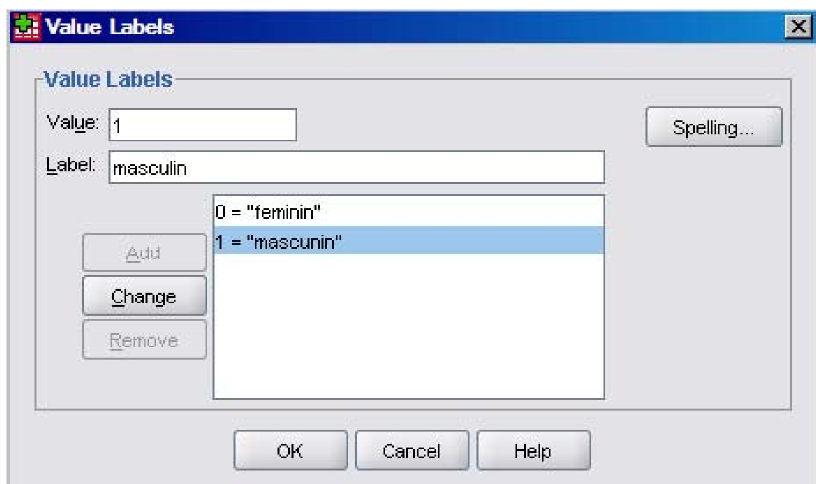
După ce am încheiat etichetarea variabilelor, trecem la etichetarea valorilor, acolo unde este cazul. Ceea ce discutăm acum se va finaliza prin introducerea unor informaţii în coloana **Values** din **Variable View**. Am precizat că nu tot timpul este nevoie să etichetăm valorile. Unele variabile au variante de răspuns care nu mai necesită explicaţii. Vârsta este măsurată în ani împliniţi : ştim ce înseamnă 46. Salariul din luna trecută este măsurat în lei : ştim ce înseamnă 1.350. Numărul de camere pe care îl are locuinţa este... un număr : ştim ce înseamnă 2. În schimb, alte variabile nu sunt măsurate la fel de intuitiv. Variantele de răspuns sunt exprimate numeric prin atribuirea unor coduri. Fericirea este măsurată prin întrebarea : „Luând în considerare toate aspectele vieţii dvs., aţi spune că sunteţi... 1. Foarte fericit, 2. Destul de fericit, 3. Nu prea fericit, 4. Deloc fericit". Respondentul alege o etichetă, dar în baza de date introducem codul. Dacă nu etichetăm codul, nu ştim ce reprezintă acesta. Pentru precizarea ocupaţiei, respondentul trebuie să aleagă dintre mai multe variante de răspuns : fiecare are un cod. Fiecare cod trebuie etichetat. Putem eticheta codurile fie în **Variable View**, fie în sintaxă, folosind o comandă simplă. Să ne ocupăm de prima variantă. Dăm click pe celula din dreptul variabilei dorite, aici d1, şi al coloanei **Values**. Se vor activa trei puncte pe care dăm click. Se va deschide fereastra din figura 4.4a.

Figura 4.4. Variable View : etichetarea valorilor variabilei

(a)



(b)



Toate secţiunile sunt iniţial goale. În celula **Value** introducem codul 0, iar în celula **Label** vom introduce eticheta „feminin". Se va activa butonul **Add**, pe care îl apăsăm. Continuăm cu codul 1 : în celula **Value**, introducem codul 1, iar în celula **Label** vom introduce eticheta „masculin". Am putea avea şi un cod de nonrăspuns. Pentru a verifica acest lucru trebuie să realizăm un tabel de frecvenţă pentru variabila d1 folosind meniul **Analyze > Descriptive statistics > Frequencies**. Dacă ar exista un cod de nonrăspuns, atunci ar trebui să îi alocăm şi acestuia o etichetă. Dacă am tastat greşit, de pildă, „masculin", atunci vom

selecta în celula mare eticheta scrisă greşit, vom modifica în celula **Label** şi vom apăsa butonul **Change** (figura 4.4b).

Este mai rapid să utilizăm sintaxa, care este la fel de simplă ca cea utilizată la etichetarea variabilelor : **VALUE LABELS** sau **VAL LAB**. Pentru etichetarea variabilelor folosite ca exemplu aici, sintaxa va fi :

VALUE LABELS d1
0 „feminin”
1 „masculin”
VALUE LABELS d2
1 „agricultor”
2 „muncitor (meserias)”
3 „tehnician, maistru, functionar”
4 „ocupatie cu studii superioare”
5 „alta ocupatie”
6 „elev, student”
7 „pensionar”
8 „casnica”
9 „acum sunt somer”
10 „patron”
VALUE LABEL d4
1 „salariat”
2 „pe cont propriu”
3 „patron”
4 „zilier”
VALUE LABEL d5
1 „agricultura”
2 „industrie, constructii”
3 „transporturi, telecomunicatii”
4 „comerţ, turism, intermedieri etc.”
5 „invatamant, cultura, cercetare, proiectare”
6 „sanatate”
7 „altele”

Structura sintaxei este aceeaşi ca la **VARIABLE LABELS**, cu diferenţa că etichetele şi codurile sunt trecute pe rânduri separate.

În chestionar, la d5, varianta de răspuns cu codul 4 are o etichetă ceva mai lungă : „comerţ, turism, intermedieri (financiare, imobiliare, pariuri etc.)”. SPSS permite un număr limitat de caractere pentru etichetele valorilor, de aceea am preferat să folosesc „etc.” în locul informaţiei dintre paranteze. Dacă nu aş fi trunchiat eticheta, ar fi făcut-o SPSS, numai că într-un mod mai puţin intuitiv de citit. Aflaţi care este numărul maxim de caractere pe care le permite SPSS pentru etichetele valorilor.

În chestionarul DCV 2010 există mai multe variabile care au aceleaşi variante de răspuns, deci aceleaşi etichete. Putem folosi o singură comandă de etichetare a valorilor acestor variabile. Să luăm, de exemplu, variabilele d14-d27. Întrebarea din chestionar este : „În viaţa fiecăruia intervin o mulţime de condiţii şi împrejurări. Ele pot fi mai bune sau mai puţin bune. Mai jos sunt menţionate o serie de asemenea aspecte. Vă rugăm să le caracterizaţi, în ceea ce vă priveşte, încercuind cifra corespunzătoare. Alegeţi un singur răspuns la fiecare întrebare”. Variabilele cărora li se aplică această întrebare sunt :

		Foarte proastă(e)	Proastă(e)	Satisfă-cătoare	Bună(e)	Foarte bună(e)	Nu e cazul
D14	Sănătatea dvs.	1	2	3	4	5	–
D15	Relaţiile din familie	1	2	3	4	5	98
D16	Locuinţa dvs.	1	2	3	4	5	–
…		1	2	3	4	5	
D27	Posibilităţile existente de petrecere a timpului liber (de recreere)	1	2	3	4	5	–

Sintaxa de etichetare a valorilor acestor variabile va fi :

VALUE LABELS d14 d15 d16 d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 d27

1 „foarte proasta(e)”

2 „proasta(e)”

3 „satisfăcătoare”

4 „buna(e)”

5 „foarte buna(e)”

98 „nu e cazul”

Nu trebuie decât să notăm după **VALUE LABELS** lista de variabile care au aceleaşi variante de răspuns şi aceleaşi etichete ale variantelor de răspuns. Simplu, nu ? Şi mult mai rapid decât dacă am fi utilizat interfaţa grafică.

Aţi observat, probabil, că nu folosesc diacritice în etichete. În această carte, în tabelele copiate din SPSS, am preferat să nu folosesc diacritice pentru a reproduce cât mai fidel senzaţia din timpul interacţiunii dvs. cu programul. Nu toţi utilizatorii au computerele setate pentru a recunoaşte diacriticile. De aceea, pentru a avea compatibilitate pe toate computerele, prefer să nu le utilizez. Cea mai neplăcută situaţie ar fi ca programul să nu le recunoască şi să le înlocuiască cu un semn de întrebare sau cu un alt caracter. De asemenea, conform manualului programului, utilizarea diacriticelor poate creşte considerabil dimensiunea bazei de date, ceea ce duce la creşterea timpului de deschidere a fişierului şi de rulare a analizelor.

4.2. Definirea nonrăspunsurilor

Nonrăspunsurile (**missing values**) reprezintă absenţa răspunsului valid. Nonrăspunsurile pot fi clasificate în două tipuri generale : cele care ţin de aplicarea chestionarului ca întreg persoanelor care ar trebui selectate conform schemei de eşantionare (*unit nonresponse*) şi cele care ţin de absenţa răspunsurilor la anumite întrebări din chestionar în cazul unei persoane selectate în eşantion (*item nonresponse*).

Primul tip de problemă apare, de exemplu, din cauza cadrelor de eşantionare care nu sunt actualizate sistematic, cum ar fi lista persoanelor cu drept de vot. Astfel, operatorul, când vizitează adresa primită, s-ar putea să nu mai găsească persoana inclusă în eşantion pentru că aceasta s-a mutat, a decedat etc. O altă cauză a nonrăspunsului de acest gen îl reprezintă dificultatea tot mai mare de a-i convinge pe oameni să răspundă la chestionare : aceştia nu au încredere în operatori, s-au plictisit din cauza solicitărilor frecvente primite de la diferite instituţii care realizează astfel de cercetări, nu au încredere în modul cum sunt gestionate răspunsurile pe care le oferă etc. O analiză detaliată a acestor probleme este realizată de Ineke Stoop (2005) în lucrarea sa intitulată sugestiv *The Hunt for the Last Respondent*. Tot în această direcţie a existat şi există o preocupare constantă în diferite anchete comparative, cum ar fi *European Social Survey*, care oferă acces la o documentaţie vastă în această zonă şi nu numai.

Al doilea tip de problemă apare, de exemplu, din cauza neatenţiei operatorului care sare peste o întrebare, refuzului de a răspunde al persoanei intervievate, care consideră întrebarea prea personală, modului cum a fost formulată o întrebare astfel încât respondentul care nu deţine informaţia respectivă se vede nevoit să declare că nu ştie răspunsul etc. În principiu, acest gen de nonrăspuns poate fi evitat prin modul cum sunt formulate întrebările şi prin pregătirea riguroasă a operatorilor de teren. Însă, în realitate, multe chestionare conţin răspunsuri de tip „nu ştiu" sau „nu răspund". Acestea nu sunt răspunsuri valide şi trebuie tratate separat în baza de date. Şi nonrăspunsul este însă un fel de răspuns, aşa că, privind din perspectiva metodologului, ar fi util să realizăm un profil al acestor persoane pentru ca în cercetarea următoare să minimizăm aceste probleme.

În această secţiune, mă voi referi doar la al doilea tip de nonrăspuns (*item nonresponse*). Nu voi analiza problemele care îl generează, ci doar cum putem lucra în SPSS cu acest gen de date. În SPSS, ca şi în alte programe de statistică, de altfel, nonrăspunsul este denumit *missing value*. De multe ori, în practică, am întâlnit mai frecvent denumirea în limba engleză, şi nu cea în limba română. Acesta este doar rezultatul utilizării frecvente de către cercetători a programelor de analiză a datelor care au interfaţa în limba engleză. În bazele de date se folosesc coduri speciale pentru nonrăspunsuri. Cel mai adesea am întâlnit codurile 97, 98 şi/ sau 99, respectiv derivate ale acestora : 7, 997, 9997, 8, 998, 9998, 9, 999, 9999

etc. În bazele de date internaţionale se folosesc (şi) alte coduri : –5, –4, –3, –2, –1. Nu trebuie să le folosiţi numai pe acestea. Important este să utilizăm un cod pentru nonrăspuns care este foarte diferit de răspunsul valid. Să luăm câteva exemple :

- Variabila d4 din DCV 2010, „statutul ocupaţional”, are patru răspunsuri valide : salariat (codul 1), pe cont propriu (codul 2), patron (codul 3) şi zilier (codul 4). În mod normal, ar trebui să primim răspunsuri valide de la toţi respondenţii pentru că este o întrebare uşor de înţeles, cu variante clare. S-ar putea însă ca un respondent să nu dorească să declare statutul său ocupaţional curent. Acest nonrăspuns va fi codificat cu 9, 99, 999 sau orice altă valoare similară sau putem utiliza codul –2, similar cercetării WVS 2012.
- Variabila d30 din DCV 2010, „Cum apreciaţi serviciul de pensii din România ? ”, are cinci răspunsuri valide : foarte prost (codul 1), prost (codul 2), satisfăcător (codul 3), bun (codul 4) şi foarte bun (codul 5). Întrebarea este aplicată tuturor respondenţilor. Un respondent care nu are pensie sau nu cunoaşte pe cineva care are pensie s-ar putea să declare că nu ştie să evalueze acest sistem. Acest nonrăspuns va fi codificat cu 8, 98, 998 sau orice altă valoare similară sau putem utiliza codul –1 similar cercetării WVS 2012.
- Variabila d10 din DCV 2010, „starea civilă”, are şase răspunsuri valide : necăsătorit (nu a fost căsătorit niciodată) (codul 1), căsătorit (codul 2), divorţat (codul 3), separat (codul 4), văduv (codul 5) şi altă situaţie (codul 6). Respondenţii care aleg codurile 1, 3, 4, 5 sau 6 sunt rugaţi să răspundă la o întrebare suplimentară : „Aveţi un partener de viaţă (cu care locuiţi împreună, aveţi menaj comun) ? ”. Cei care au răspuns codul 2, adică sunt căsătoriţi, nu mai trebuie să răspundă la această întrebare. Nu li se aplică. Acesta este un tip aparte de nonrăspuns, denumit „nu e cazul”, care va fi codificat cu 7, 97, 997 sau orice altă valoare similară sau putem utiliza codul –3 similar cercetării WVS 2012.

În Access sau în programul pe care îl utilizăm pentru introducerea datelor, am definit deja aceste nonrăspunsuri pentru a uşura procesul de introducere a datelor şi de curăţare a bazei de date. Teoretic, nu ar trebui să mai introducem coduri în faza de curăţare.

Unii cercetători preferă să nu instruiască SPSS că „nu ştiu” (98), „nu răspund” (99) sau „nu e cazul” (97) sunt nonrăspunsuri, lăsând acest lucru pentru momentul analizei pe care o va face. Alţii preferă ca baza dată echipei de cercetare sau altor utilizatori să aibă deja nonrăspunsurile definite.

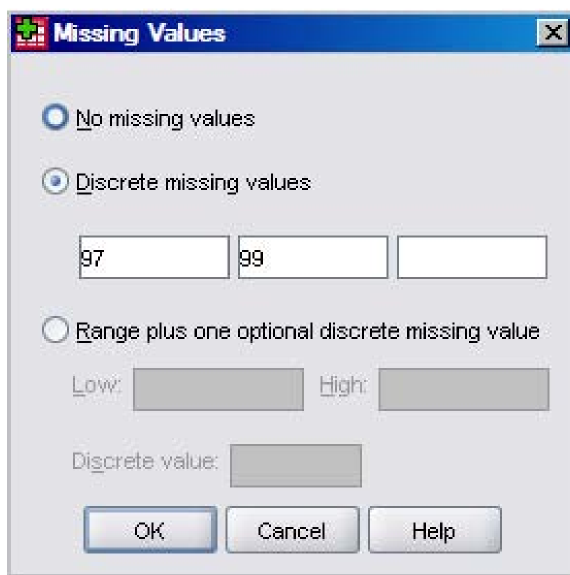
Nonrăspunsurile pot fi definite fie în **Variable View**, fie în sintaxă. Vă recomand a doua variantă.

În **Variable View**, mergeţi cu cursorul pe celula din dreptul variabilei care vă interesează (pe rând) şi al coloanei **Missing** (pe coloană). Se vor activa cele trei puncte, None ..., pe care dăm click. După ce am realizat modificările, în locul

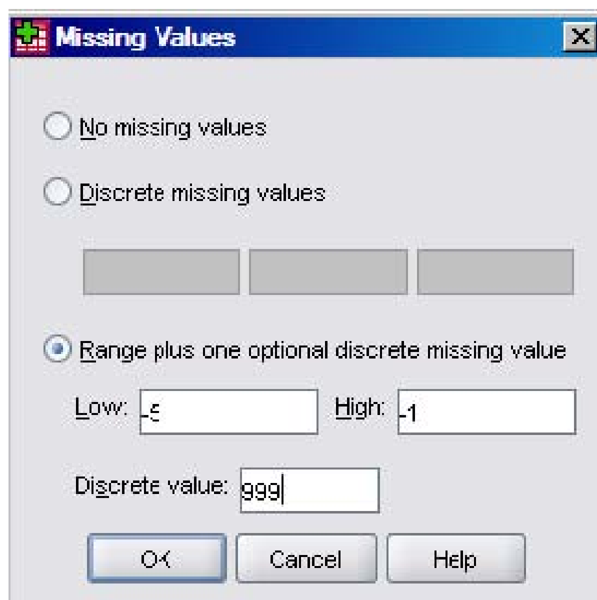
cuvântului **None**, vom observa valorile pe care le-am definit ca nonrăspunsuri. Se deschide fereastra din figura 4.5. În această fereastră, iniţial, este selectată opţiunea **No missing values**.

Figura 4.5. Definirea nonrăspunsurilor în Variable View

(a)



(b)



Există două opţiuni pe care le putem bifa.

Opţiunea **Discrete missing values** este folosită atunci când variabila are maximum trei coduri de nonrăspuns, adică maximum trei tipuri de nonrăspuns : „nu ştiu", „nu răspund", „nu e cazul". În figura 4.5a definim nonrăspunsurile pentru variabila d3, din DCV 2010, care conţine informaţii despre anul de când respondentul are ocupaţia pe care a declarat-o la d2. Prin realizarea unui tabel de frecvenţă, am observat că d3 are două tipuri de nonrăspuns : „nu ştiu/nu răspund", care a primit codul 99, şi „nu e cazul", care a primit codul 97. 463 de respondenţi au primit codul 97 pentru că nu au o ocupaţie (sunt inactivi pe piaţa muncii), iar 165 nu au ştiut sau nu au dorit să precizeze anul de când au ocupaţia actuală.

Opţiunea **Range plus one optional discrete missing values** este folosită atunci când avem mai mult de trei tipuri de nonrăspuns : „nu ştiu", „nu răspund", „nu e cazul", „întrebarea nu a fost adresată în anul respectiv" etc. De exemplu, în WVS 2012 avem coduri de la –5 la –1 : „missing : unknown" (codul –5), „not asked in survey" (–4), „not applicable" (–3), „no answer" (–2), „don't know" (–1). La **Low** introducem –5, la *high* introducem –1. În figura 4.5b definim nonrăspunsurile pentru o variabilă care are coduri de nonrăspuns de la –5 la –1, dar şi un cod 999. În principiu, această opţiune acoperă toate situaţiile posibile.

De unde ştim ce coduri trebuie să introducem în aceste celule ? Am precizat deja că realizăm un tabel de frecvenţă pentru fiecare variabilă pentru care vrem să definim nonrăspunsurile. Să luăm ca exemplu variabila d15, „Cât de mulţumit sunteţi de relaţiile din familie ? " (tabelul 4.1).

Tabelul 4.1. Tabel de frecvenţă înainte de definirea nonrăspunsurilor

d15 relatiile din familie					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 foarta proasta(e)	8	.7	.7	.7
	2 proasta(e)	14	1.2	1.2	1.9
	3 satisfacatoare	107	9.2	9.2	11.1
	4 buna(e)	666	57.4	57.4	68.5
	5 foarte buna(e)	289	24.9	24.9	93.4
	98 nu e cazul	42	3.6	3.6	97.0
	99 nu stiu/nu raspund	35	3.0	3.0	100.0
	Total	1161	100.0	100.0	

Tabelul este realizat după ce am încheiat etapa de etichetare a variabilelor şi a valorilor variabilelor. Observăm că această variabilă are cinci răspunsuri valide : „foarte proastă(e)" (codul 1), „proastă(e)" (codul 2), „satisfăcătoare" (codul 3), „bună(e)" (codul 4) şi „foarte bună(e)" (codul 5). De asemenea, are două tipuri de nonrăspuns : „nu e cazul", codul 98, şi „nu ştiu/nu răspund", cumulate în codul 99. Observaţi că nu există o regulă strictă care impune utilizarea aceloraşi coduri

în toate cercetările. Trebuie doar să existe o anumită consistenţă pentru a face mai uşoară tranziţia de la o cercetare la alta. Cei care au răspuns „nu e cazul” s-au gândit, probabil, că întrebarea se referă la o relaţie maritală de tip soţ-soţie. Dacă inspectăm această ipoteză, observăm că toţi cei 42 de respondenţi din categoria „nu e cazul” (98) sunt necăsătoriţi, divorţaţi sau văduvi. Aşadar, ei au considerat că nu pot răspunde la această întrebare. În mod normal, în faza de curăţare, dacă cercetătorului i se pare ciudat ca o persoană să nu răspundă la o întrebare, ar putea verifica chestionarele pentru o posibilă eroare de introducere sau ar putea chiar discuta cu operatorul de teren solicitând, uneori, refacerea chestionarului. Revenind la definirea nonrăspunsurilor, am aflat că trebuie să introducem în celulele **Discrete missing values** codurile 98 şi 99.

Putem automatiza activitatea de definire a nonrăspunsurilor folosind sintaxa. Comanda este la fel de simplă ca celelalte două comenzi discutate : **VARIABLE LABELS** şi **VALUE LABELS**. Comanda pentru nonrăspunsuri este **MISSING VALUES**. Mai exact, pentru variabila d15 comanda este :

MISSING VALUES d15 (98, 99).

La fel ca la **VALUE LABELS**, putem utiliza aceeaşi linie de comandă pentru mai multe variabile care au coduri similare la nonrăspunsuri. De exemplu, succesiunea de variabile d15-d27 se află în această situaţie. Aşadar, comanda va arăta astfel :

MISSING VALUES d15 d16 d17 d18 d19 d20 d21 d22 d23 d24 d25 d26 d27 (98, 99).

Observăm cât de uşor putem defini nonrăspunsurile în acest meniu faţă de meniul **Variable View**, unde am fi dat mai multe clickuri pentru fiecare variabilă în parte. În plus, oricând dorim, putem consulta sintaxa, reamintindu-ne ce am lucrat sau pentru a o rula din nou pe o bază „curată”.

Pentru că orice proces de învăţare presupune căutare de informaţie, vă invit să aflaţi ce element din comanda **MISSING VALUES** puteţi şterge fără a afecta rezultatul final.

4.3. Verificarea introducerii eronate a unor coduri

Dacă am folosit un program de introducere a datelor care restricţionează operatorul de introducere să introducă greşit o valoare în afara amplitudinii răspunsurilor posibile, în principiu, atunci putem sări această etapă, deşi niciodată nu strică o verificare.

Verificarea este o operaţie simplă care presupune doar inspectarea tabelelor de frecvenţă pentru fiecare variabilă din baza de date. Deja am precizat că aceste tabele se realizează din meniul **Analyze > Descriptive statistics > Frequencies**.

Trebuie să comparăm ce apare în tabel cu ceea ce este scris în chestionar. Această operaţiune poate fi realizată fie înaintea definirii nonrăspunsurilor, fie ulterior. Este util, în schimb, să fie încheiată operaţiunea de etichetare a valorilor variabilelor pentru a vedea ce reprezintă fiecare cod.

4.4. Validarea logică prin urmărirea filtrelor şi a unor întrebări factuale

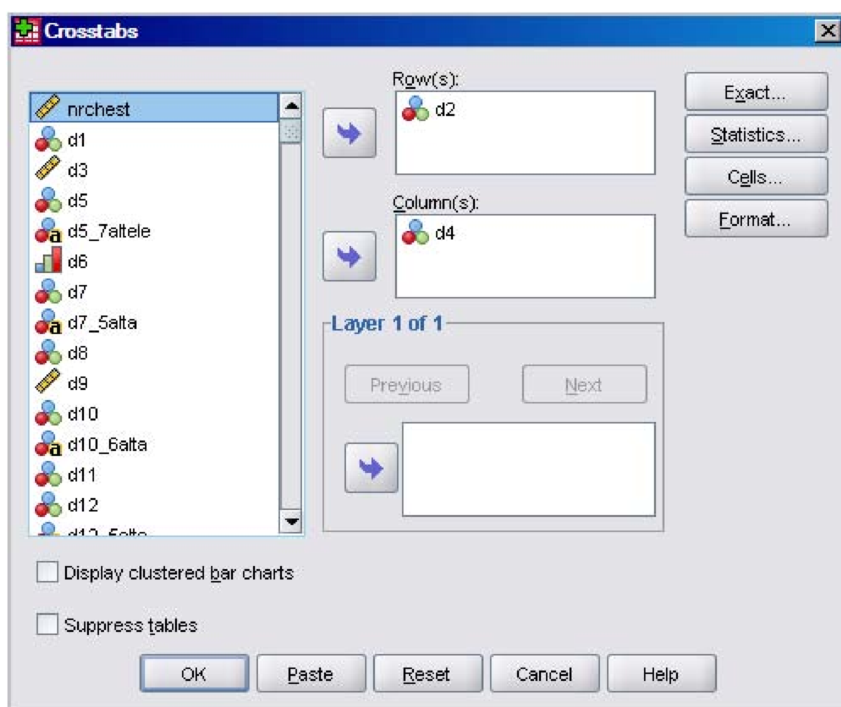
Cei care au declarat că sunt elevi (codul 6), pensionari (codul 7), casnici (codul 8) sau şomeri (codul 9) la variabila d2, ocupaţia principală actuală, trebuiau să răspundă apoi direct la întrebarea d6. Aşadar, aceştia trebuie să aibă, la întrebările dintre d2 şi d6, un cod de nonrăspuns de tipul „nu e cazul". Pur şi simplu, întrebările dintre d2 şi d6 nu li se aplică. Dacă am lucrat corect în programul de introducere a datelor sau în alt program similar nu ar trebui să apară erori de introducere. Putem verifica simplu dacă filtrul a fost respectat, realizând un tabel de contingenţă folosind meniul **Analyze > Descriptive statistics > Crosstabs** între d2 şi fiecare dintre întrebările de până la d6. În tabelul 4.2 este prezentat un exemplu de încrucişare între d2 (ocupaţia principală actuală) şi d3 (statutul ocupaţional). Conform chestionarului, în celulele rezultate din intersecţia dintre rândurile ce conţin codurile 6, 7, 8 şi 9 şi coloanele date de răspunsurile valide la d4 şi codul 99 („nu ştiu/nu răspund") ar trebui să apară valoarea 0, adică nici o persoană. Observăm că aici filtrul este respectat: apar persoane doar la intersecţia dintre codurile 6-9 la d2 şi codul 97 („nu e cazul") la d4.

Tabelul 4.2. Tabel de contingenţă ce verifică un filtru, dar este folosit pentru validare logică (1)

d2 ocupatia dvs. actuala (principala) * d4 statutul ocupational Crosstabulation								
Count								
		d4 statutul ocupational						Total
		1 salariat	2 pe cont propriu	3 patron	4 zilier	97	99	
d2 ocupatia dvs. actuala (principala)	1 agricultor	2	228	0	29	0	0	259
	2 muncitor (meserias)	215	14	1	12	0	5	247
	3 tehnician, maistru, functionar	72	2	0	0	0	0	74
	4 ocupatie cu studii superioare	98	8	0	0	0	0	106
	6 elev, student	0	0	0	0	52	0	52
	7 pensionar	0	0	0	0	267	0	267
	8 casnic	0	0	0	0	62	0	62
	9 acum sunt somer	0	0	0	0	82	0	82
	10 patron	0	0	12	0	0	0	12
Total		387	252	13	41	463	5	1161

Revenim la modul în care am realizat tabelul de contingenţă. Deşi acestui subiect i se dedică o secţiune specială, cred că este util să vedem paşii elementari în realizarea acestui tip de tabel şi aici. Accesând meniul **Analyze > Descriptive statistics > Crosstabs** se deschide fereastra din figura 4.6. Un tabel de contingenţă are două variabile. O variabilă, prin categoriile ei, dă rândurile tabelului. Cealaltă variabilă, prin categoriile ei, dă coloanele tabelului. Fiecare celulă din tabel ne arată numărul persoanelor care se regăsesc în două categorii simultan : 2 persoane sunt agricultori salariaţi, 215 sunt muncitori salariaţi, 8 persoane au o ocupaţie care necesită studii superioare şi lucrează pe cont propriu etc. Care celulă din tabel prezintă o informaţie inconsistentă ? Cum puteţi explica această inconsistenţă şi ce ar trebui să faceţi pentru a o corecta ?

Figura 4.6. Crosstabs : realizarea unui tabel de contingenţă



Prefer să introduc în rând (celula **Row**) variabila care are cele mai multe variante de răspuns pentru a rezulta un tabel pe verticală, uşor de inserat într-o pagină A4 orientată portret. Pe coloană (celula **Column**) introduc cealaltă variabilă. Pentru ce avem nevoie acum este suficient să apăsăm **OK**. Va rezulta tabelul cu frecvenţele absolute, adică cu celulele arătând numărul de agricultori care sunt salariaţi, numărul de agricultori care lucrează pe cont propriu, numărul de muncitori care sunt salariaţi ş.a.m.d.

Aţi reuşit să observaţi inconsistenţa ? O persoană care a declarat că ocupaţia sa este muncitor a indicat ca statut ocupaţional faptul că este patron. Aici avem de-a face cu o validare logică a informaţiei conţinute de baza de date. Cei care au ales varianta 2 la d2 puteau răspunde la d4, deci nu avem un filtru. În schimb, logic ar fi ca un muncitor să nu se declare patron. Primul lucru pe care ar trebui să îl facem este să consultăm chestionarul completat. Dacă nu găsim răspunsul la această inconsistenţă în el, atunci va trebui să discutăm cu operatorul de teren pentru a ne clarifica situaţia.

Să presupunem că am identificat de unde vine această problemă. Am aflat că, de fapt, persoana respectivă are ocupaţia de patron. Deci a fost introdus greşit în baza de date codul 2 în loc de codul 10 la d2. Aşadar trebuie să modificăm informaţia în baza de date. Acest lucru se face prin recodificarea variabilei d2. Dar pentru a face recodificarea trebuie să aflăm id-ul unic al acelui respondent pentru a fi siguri că modificăm numai ce ne interesează. Acest lucru se face urmând mai mulţi paşi :

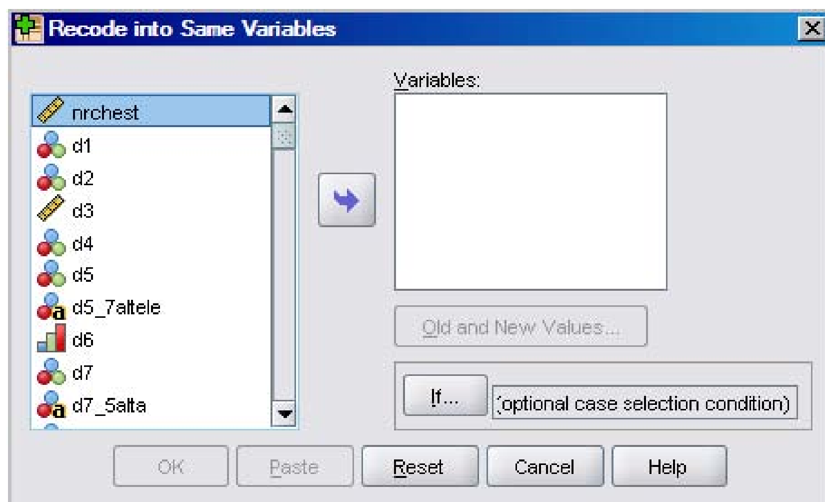
- filtrăm baza de date astfel încât să rămână activ doar cazul care are codul 2 la d2 şi codul 3 la d4, adică muncitorul care a declarat statutul patron. Filtrul este : **d2 = 2 & d4 = 3**. Mergem în **Data > Select Cases > If condition is satisfied > If > introducem condiţia > Continue > OK** ;
- realizăm un tabel de frecvenţă pentru variabilele care sunt folosite pentru condiţie şi pentru variabila care conţine id-ul unic, aici nrchest ;
- verificăm dacă filtrul activ este cel dorit ;
- citim tabelul de frecvenţă pentru variabila nrchest şi aflăm că acel caz are id-ul 312 ;
- dezactivăm filtrul.

Acum putem trece la recodificarea variabilei d2, pentru că ea conţine eroarea. Recodificarea se va face din meniul **Transform > Recode into Same Variables**. Acest meniu va înlocui, pentru cazul cu id-ul 312, codul 2 cu codul 10. Figura 4.7 prezintă etapele acestui proces :

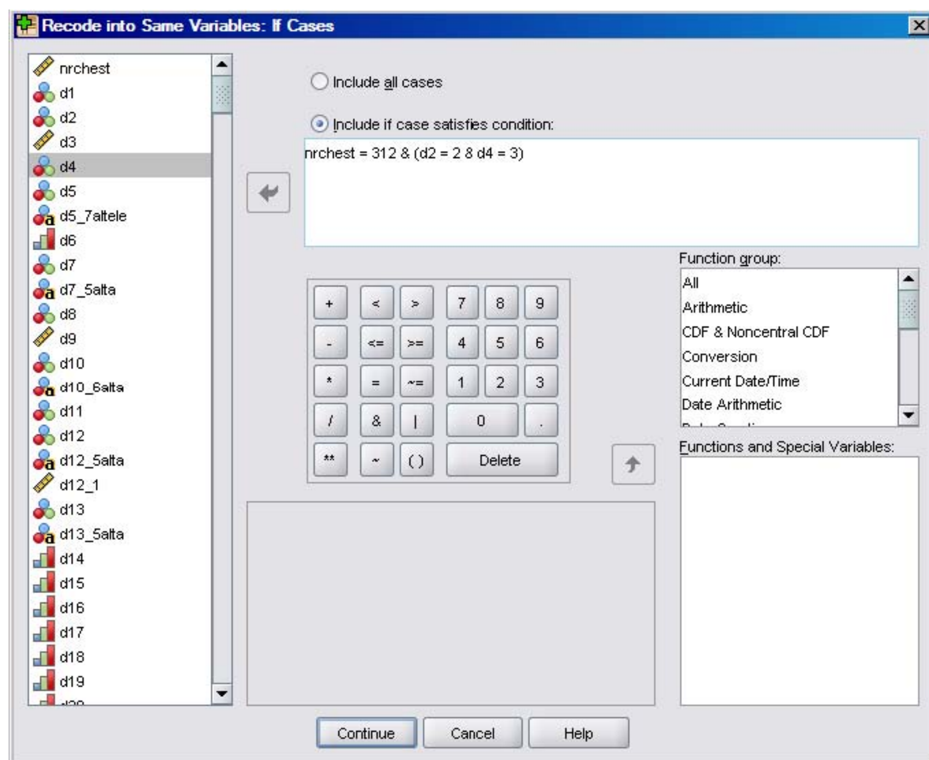
- Apăsând butonul **If > Include if case satisfies condition**, punem condiţia ca modificarea să fie realizată doar pentru cazul cu id-ul 312. Aici am notat şi că **d2 = 2 & d4 = 3**. Apăsăm **Continue** (figura 4.7b).
- Apăsăm butonul **Old and New Values**.
- Înlocuim codul 2 (**Old Value > Value**) cu codul 10 (**New Value > Value**). Apăsăm butonul **Add**, apoi apăsăm butonul **Continue** (figura 4.7c).
- Am revenit în fereastra iniţială, unde apăsăm butonul **OK**.
- Refacem tabelul de contingenţă dintre d2 şi d4, pentru a verifica dacă modificarea s-a făcut conform aşteptărilor (tabelul 4.3).

Figura 4.7. Recode into Same Variables

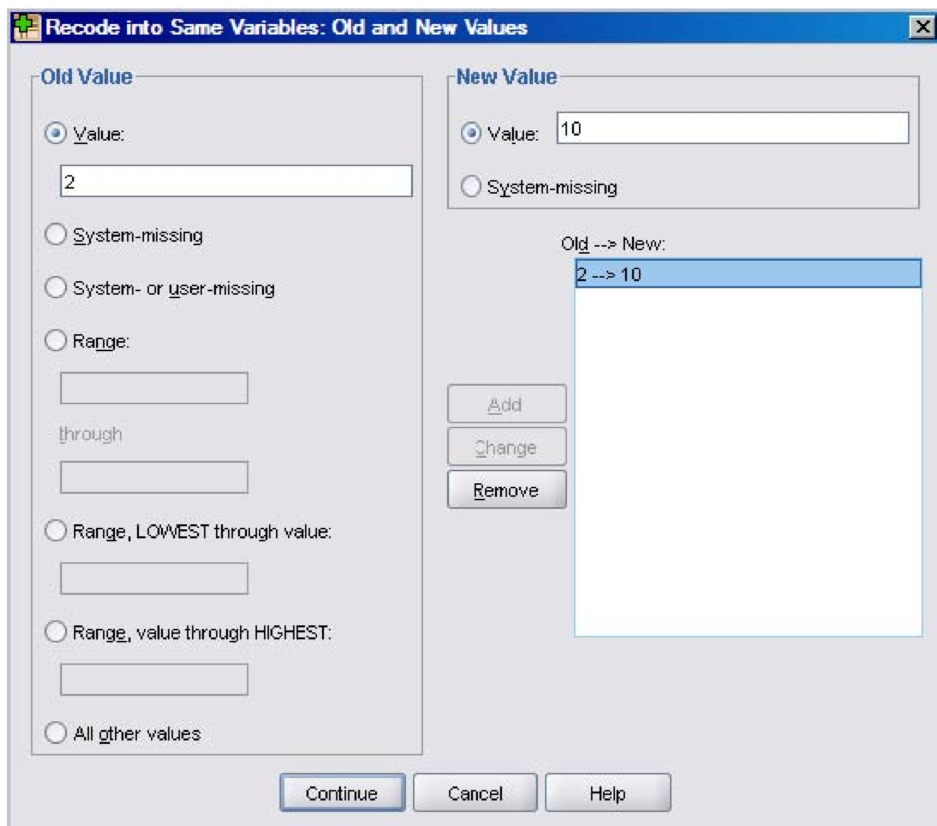
(a)



(b)



(c)



Utilizarea condiţiilor în meniul **Recode into Same Variables** nu este obligatorie. Acest lucru a fost impus de situaţia discutată. De multe ori însă folosim doar comenzile activate de butonul **Old and New Values**. Fereastra care se deschide are mai multe secţiuni (figura 4.7b). În secţiunea **Old Value** introducem valorile iniţiale : cele pe care dorim să le recodificăm. În secţiunea **New Value** introducem valorile noi : cele în care vor fi recodificate valorile iniţiale. În secţiunea **Old --> New**, după ce apăsăm butonul **Add**, apar modificările pe care dorim să le facem. Aici am dorit să modificăm doar un cod : 2 în 10. De aceea am folosit **Old Value > Value**. Dacă am fi vrut să modificăm o serie de numere consecutive, să zicem 2-6 în 10, atunci am fi folosit **Old Value > Range 2 through 6**. Dacă am fi vrut să modificăm o serie de numere consecutive de la cea mai mică valoare la o valoare anume, să zicem de la 1 (valoarea minimă) la 4 (valoarea specifică), atunci am fi folosit **Old Value > Range, LOWEST throug value : 4**. Dacă am fi vrut să modificăm o serie de numere consecutive de la o valoare specifică la cea mai mare din serie, să zicem de la 4 (valoarea specifică) la 10 (valoarea maximă), atunci am fi folosit **Old Value > Range, value through**

HIGHEST: 10. Vom discuta despre recodificare în altă secţiune a lucrării. Ce trebuie reţinut aici este că folosirea meniului **Recode into Same Variables** suprascrie informaţia iniţială. Deci atenţie la ce modificări doriţi să faceţi.

Tabelul 4.3. Tabel de contingenţă care verifică un filtru, dar este folosit şi pentru validare logică (2)

d2 ocupatia dvs. actuala (principala) * d4 statutul ocupational Crosstabulation								
Count								
		d4 statutul occupational						Total
		1 salariat	2 pe cont propriu	3 patron	4 zilier	97	99	
d2 ocupatia dvs. actuala (principala)	1 agricultor	2	228	0	29	0	0	259
	2 muncitor (meserias)	215	14	0	12	0	5	246
	3 tehnician, maistru, functionar	72	2	0	0	0	0	74
	4 ocupatie cu studii superioare	98	8	0	0	0	0	106
	6 elev, student	0	0	0	0	52	0	52
	7 pensionar	0	0	0	0	267	0	267
	8 casnic	0	0	0	0	62	0	62
	9 acum sunt somer	0	0	0	0	82	0	82
	10 patron	0	0	13	0	0	0	13
Total		387	252	13	41	463	5	1161

Aşadar, pe lângă verificarea filtrelor, realizăm şi validarea logică prin încrucişarea unor variabile factuale. O variabilă factuală culege informaţii concrete care nu ţin de valori, atitudini, opinii, credinţe, evaluări. De exemplu, sexul sau vârsta respondentului sunt variabile factuale. Tot variabile factuale sunt salariul măsurat într-o unitate monetară, suprafaţa locuinţei în metri pătraţi, numărul de copii etc. În chestionare, din cauza neatenţiei operatorului sau poate chiar dintr-o scăpare de design a cercetătorului, se mai întâmplă ca o persoană să declare ceva la o variabilă factuală, acel ceva fiind incompatibil cu ce declară la altă variabilă factuală aflată într-o relaţie logică cu cea dintâi. Un bărbat nu are voie să răspundă la întrebarea „Aţi făcut vreodată avort ?”. În schimb, are voie să răspundă la întrebarea „Partenera dvs. de viaţă a făcut vreodată avort ?”. Cel care a declarat că nu suferă de vreo boală nu are voie să răspundă la întrebarea „Suferiţi de o boală cronică ?”. Dincolo de cele două tipuri de erori enunţate mai există o situaţie care, într-un fel, ţine de designul chestionarului, deci este o problemă a cercetătorului. Realitatea din teren s-ar putea să fie mai complexă decât cea pe care o

cunoaşte sau şi-o imaginează cercetătorul. De exemplu, un cercetător s-ar putea aştepta ca o persoană care declară că este pensionar să nu mai ofere un răspuns valid la rubrica „Vă rugăm să ne spuneţi care a fost suma de bani încasată luna trecută din salarii”, ci doar la rubrica „Vă rugăm să ne spuneţi care a fost suma de bani încasată luna trecută din pensii”. Salariul, teoretic, este specific unei persoane angajate formal, cu contract de muncă. Totuşi, salariul poate fi atribuit şi persoanelor care nu sunt angajate formal, ci prestează diferite servicii informal („la negru”). Când ne gândim la un pensionar ne imaginăm că salariul acestuia este pensia, deci nu mai prestează servicii, cel puţin formalizate. Nivelul de trai redus din România şi, implicit, al pensiilor îi determină pe mulţi pensionari să lucreze informal. De exemplu, un pensionar se poate „angaja” ca paznic de noapte la o firmă. Acesta primeşte o pensie, dar şi un salariu, chiar dacă acel salariu nu este înregistrat legal. Aici intervine altă problemă : să presupunem că cercetătorul admite că acest gen de situaţii este veridic, astfel încât îi va adresa întrebarea referitoare la salariu şi pensionarului. Pensionarul, în schimb, fiind conştient că salariul său nu este înregistrat legal, s-ar putea să refuze să răspundă la întrebarea legată de salariu şi să accepte să răspundă doar la întrebarea legată de pensie. Astfel apare nonrăspunsul şi, implicit, discuţia se mută în zona de distorsiune a realităţii, de modificare a reprezentativăţii eşantionului.

Revenind la problema validării logice prin încrucişarea variabilelor factuale, primul pas ce trebuie făcut este să identificăm în chestionar toate interacţiunile posibile dintre variabilele factuale. Apoi, realizând tabele de contingenţă, aşa cum am discutat deja, scanăm datele pentru erori. Termenul „eroare” este poate prea tranşant. Cercetătorii trebuie să consulte chestionarele tipărite şi, eventual, să contacteze din nou respondentul pentru clarificări. Abia apoi se intervine în baza de date. Validarea logică poate fi inclusă chiar în partea de design a cercetării şi chestionarului. De exemplu, într-un studiu prin care se dorea estimarea incidenţei consumului diferitelor tipuri de droguri, cercetătorul a introdus în lista de droguri şi unul fictiv. Dacă în chestionar apăreau răspunsuri valide la acest drog (respondentul „spunea” că a consumat, cu o anumită frecvenţă, în anumite condiţii etc.), atunci cercetătorul afla imediat că operatorul de teren nu a fost onest şi a completat el însuşi acel chestionar.

În DCV 2010, există variabila d39 : „Caracterizaţi măsura în care puteţi influenţa luarea deciziilor în organizaţia în care lucraţi” cu variantele de răspuns „foarte scăzută” (codul 1), „scăzută” (codul 2), „satisfăcătoare” (codul 3), „ridicată” (codul 4), „foarte ridicată” (codul 5). În chestionar, în dreptul acestei variabile, există şi varianta „nu e cazul” (codul 98). Logica este simplă : o persoană care nu are un loc de muncă nu poate să evalueze măsura în care are libertate de decizie acolo (organizaţie este un termen generic pentru toate locurile de muncă fie că acestea sunt în firme, instituţii publice, ONG-uri etc.). Prin urmare, trebuie să verificăm legătura logică cu variabila factuală d2, „ocupaţia dvs actuală principală”, care are 10 variante de răspuns : „agricultor” (codul 1), „muncitor (meseriaş)” (codul 2), „tehnician, maistru, funcţionar” (codul 3),

„ocupaţie cu studii superioare" (codul 4), „altă ocupaţie" (codul 5), „elev, student" (codul 6), „pensionar" (codul 7), „casnic" (codul 8), „acum sunt şomer" (codul 9) şi „patron" (codul 10). Logic ar fi ca elevii/studenţii, pensionarii, casnicii şi şomerii, adică codurile 6, 7, 8 şi 9, să nu aibă răspunsuri valide la d39. Tabelul de contingenţă de mai jos (tabelul 4.4) ne arată o încălcare a acestei logici : există 2 studenţi care evaluează libertatea de decizie ca fiind ridicată, 3 şomeri care o evaluează ca fiind foarte scăzută şi 4 şomeri care o evaluează ca fiind scăzută, 1 pensionar care o evaluează ca fiind scăzută.

Tabelul 4.4. Validare logică : tabel de contingenţă

d2 ocupatia dvs. actuala (principala) * d39 masura in care puteti influenta luarea deciziilor in organizatia in care lucrati Crosstabulation									
Count									
		d39 masura in care puteti influenta luarea deciziilor in organizatia in care lucrati							Total
		1 foarte sca-zuta	2 sca-zuta	3 satisfa-catoare	4 ridi-cata	5 foarte ridi-cata	98 nu e cazul	99	
d2 ocupatia dvs. actuala (principala)	1 agricultor	8	15	26	4	1	197	8	259
	2 muncitor (meserias)	43	66	77	29	9	18	4	246
	3 tehnician, maistru, functionar	11	18	22	18	2	3	0	74
	4 ocupatii cu studii superioare	9	21	31	23	11	11	0	106
	6 elev, student	0	0	0	2	0	50	0	52
	7 pensionar	0	1	0	0	0	266	0	267
	8 casnic	0	0	0	0	0	62	0	62
	9 acum sunt somer	3	4	0	0	0	73	2	82
	10 patron	3	0	1	1	2	5	1	13
Total		77	125	157	77	25	685	15	1161

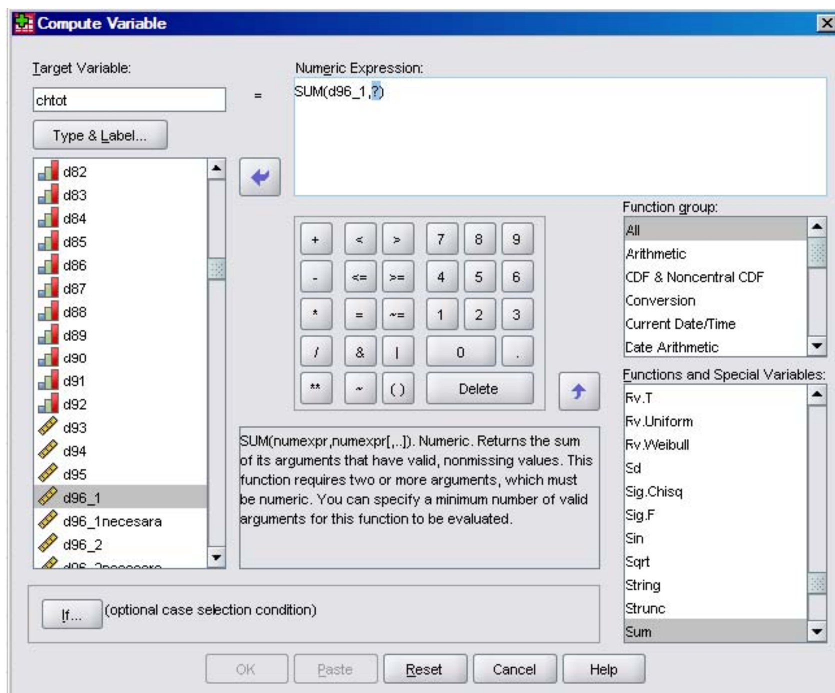
Încă o dată : modificările nu se fac automat. Am verificat chestionarele şi am constatat următoarele situaţii : (1) studenţii sunt, de fapt, persoane care au ocupaţii cu studii superioare (să nu uităm că d2 cere răspuns unic) ; (2) cei 3 şomeri sunt agricultori, iar (3) cei 4 şomeri sunt muncitori (meseriaşi). Aşadar, trebuie să recodificăm variabila d2 pentru cazurile acestea. Dacă dorim să procedăm ca mai devreme, care sunt paşii pe care trebuie să îi parcurgeţi pentru realizarea modificărilor ?

4.5. Construirea unor variabile noi

Aceasta poate fi sau nu o parte a procesului de curăţare. Unii cercetători preferă să aibă anumite variabile create încă de la început, alţii consideră că le pot construi singuri pe măsură ce este nevoie de ele. În DCV 2010, există întrebarea d96 prin care respondenţii sunt rugaţi să raporteze, în lei, sumele cheltuite pentru 9 variabile : alimente ; băuturi alcoolice, ţigări ; îmbrăcăminte, încălţăminte ; pentru locuinţă : chirie, întreţinere, reparaţii, abonamente, rate ; transport ; îngrijirea sănătăţii ; cultură, şcoală, cărţi, spectacole ; pensie alimentară ; alte cheltuieli. În unele analize, am putea fi interesaţi să lucrăm cu variabila care conţine informaţii despre cheltuielile totale ale gospodăriei în luna precedentă. Aceasta este, evident, suma tuturor acestor nouă variabile. Suma aceasta va deveni o nouă variabilă în baza de date. Fiecare respondent va avea, în dreptul său, valoarea însumată a tuturor cheltuielilor efectuate.

Pentru a calcula această sumă, folosim meniul **Transform > Compute** (figura 4.8). Decidem că numele variabilei va fi „chtot". Prefer numele scurte, formate doar din litere şi, eventual, cifre, pentru că folosesc anumite programe de statistică specializate cum ar fi HLM (Raudenbush *et al.*, 2011), care solicită aceste specificaţii. Dacă nu le respect, programul va redenumi variabilele şi, de multe ori, această operaţie automată creează nume cu care este greu de lucrat.

Figura 4.8. Crearea de noi variabile (Compute)



La **Target Variable** vom scrie numele variabilei pe care o realizăm, aici chtot. La **Numeric Expression** vom scrie formula care ne dă nouă variabilă. Putem folosi două abordări, în funcţie de necesităţi : (1) folosim funcţiile pe care le oferă SPSS, cum ar fi funcţia **SUM()**, sau (2) introducem noi o expresie de tipul a+b+c+...+n. Pentru moment, folosim funcţia **SUM()** pe care o aducem în **Numeric Expression** din secţiunea **Functions and Special Variables**. Mai întâi, dăm click în secţiunea **Function group** pe **All**, pentru a se activa funcţiile din secţiunea **Functions and Special Variables**. Apoi căutăm funcţia **SUM()** folosind acelaşi procedeu ca în lista de variabile din orice meniu ori, pur şi simplu, utilizând scroll-ul. Când o găsim, dăm dublu click pe ea şi vom vedea că va apărea în **Numeric Expression**. Iniţial, ea arată astfel : **SUM(? , ?)**. Ştergem semnele de întrebare, căutăm în lista de variabile din stânga ceea ce ne interesează, aici succesiunea d96_1-d96_9, dăm, pe rând, dublu click pe variabile sau le introducem cu săgeata în dreapta, punând virgulă între ele. Apăsăm **OK**. Putem merge în **Data View** sau **Variable View** să vedem variabila. Variabilele noi sunt create la sfârşitul bazei de date. Putem, în loc de funcţia **SUM()**, să adunăm pur şi simplu variabilele respective. Adică, în **Numeric Expression**, să fi scris : d96_1 + d96_2 + d96_3 + d96_4 + d96_5 + d96_6 + d96_7 + d96_8 + d96_9. Rezultatul este diferit şi vine din modul în care SPSS tratează nonrăspunsurile. Să presupunem că la aceste întrebări există persoane care au refuzat să răspundă sau au declarat că nu ştiu să răspundă. Aceste valori nu sunt valide şi nu vor fi luate în calcul la sumă dacă le-am definit în coloana **Missing** din **Variable View** sau folosind comanda **MISSING VALUES** în sintaxă. Dacă folosim funcţia **SUM()**, atunci variabila chtot va conţine suma variabilelor chiar dacă, la una sau mai multe dintre ele, respondentul nu a indicat o valoare validă, ci a oferit, în schimb, un nonrăspuns. Dacă folosim adunarea, atunci variabila chtot va conţine suma pentru respondenţii care au oferit răspunsuri valide la toate variabilele din şir, ştergându-i pe ceilalţi. Deci în ultima variantă o să avem mai puţine valori valide în variabila nou-creată, pentru că este folosită doar informaţia completă, pe când în prima variantă o să avem mai multe valori valide pentru că este folosită toată informaţia disponibilă. Care variantă este corectă ? Răspunsul nu este atât de evident. Aici, la cheltuieli, am putea folosi şi informaţia parţială notând totuşi în lucrarea pe care o scriem această limită a analizei. Dacă avem o variabilă latentă, cum ar fi atitudinea faţă de fumat măsurată printr-o scală compusă care conţine 5 itemi, atunci poate ar fi bine să folosim informaţia completă : altfel, scorul, adică atitudinea, s-ar baza pe o măsurătoare incompletă. Cel mai onest ar fi să construim ambele variabile, să rulăm analizele dorite cu ambele variabile separat şi să vedem dacă rezultatele se schimbă substanţial.

Variabilele nou-create de noi nu sunt etichetate automat. Va trebui să rulăm de fiecare dată sintaxele **VARIABLE LABELS** şi **VALUE LABELS**, în funcţie de nevoi.

O altă situaţie, destul de frecvent întâlnită, în care se impune folosirea meniului **Transform > Compute** se referă la calcularea vârstei. În chestionar, respondentul nu este întrebat ce vârstă are, ci în ce an s-a născut. În analize însă, suntem

interesaţi să lucrăm cu vârsta, de aceea va trebui să creăm această variabilă în baza de date. Vârsta va fi egală cu anul aplicării chestionarului minus anul naşterii. În DCV 2010, respondentul este rugat să îşi declare vârsta în ani împliniţi, situaţie care nu se aplică aici. Există însă variabila d3 care înregistrează anul din care respondentul are ocupaţia declarată la d2. În analize ne interesează să lucrăm cu variabila vechime în muncă măsurată în ani. Decidem să creăm această variabilă care se va numi „vechime". Mai întâi, trebuie să ne asigurăm că la d3 sunt definite nonrăspunsurile. Observăm că în această bază avem codul 97, aplicat celor care nu au o ocupaţie în prezent, şi codul 99, aplicat celor care nu au vrut să răspundă sau nu au ştiut unde să se încadreze în variantele puse la dispoziţie de cercetător. Dacă nu facem acest lucru, vor fi luate în considerare la calcul şi aceste coduri, noua variabilă conţinând informaţii eronate. Evident va trebui să o etichetăm pentru a şti în continuare ce reprezintă. În cadrul unei analize, lucrăm cu multe variabile şi este foarte uşor să uităm ce am făcut anterior, mai ales dacă lăsăm o pauză de câteva zile între început şi sfârşit şi lucrăm în mai multe proiecte simultan.

Posibilităţile pe care ni le oferă comanda **COMPUTE** sunt numeroase. Fiecare le va folosi pe cele care îi sunt utile în analize.

4.6. Exerciţii

Pentru aceste exerciţii utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultat(ă/e) în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a *Grupului Românesc pentru Studiul Valorilor Sociale* (http : //www.romanianvalues.ro).

1. Deschideţi baza de date finală creată la exerciţiul 13 din capitolul 2. Definiţi proprietăţile variabilelor din baza de date în Variable View.
2. Realizaţi câte un tabel de frecvenţă pentru fiecare variabilă din baza de date. Există coduri introduse eronat ? Dacă da, cum explicaţi această greşeală ?
3. Identificaţi în cele patru pagini de chestionar alese în exerciţiile din capitolul 2 întrebările filtru. Verificaţi dacă filtrele au fost respectate.
4. Identificaţi, în cele patru pagini de chestionar, variabile care pot fi folosite în procesul de validare logică. Verificaţi dacă există situaţii în care logica a fost încălcată.
5. Deschideţi baza de date WVS 2012. Realizaţi câte un tabel de frecvenţă pentru fiecare variabilă din chestionar. Există coduri introduse eronat ? Dacă da, cum explicaţi această greşeală ?
6. Identificaţi în chestionarul WVS 2012 întrebările filtru. Verificaţi dacă filtrele au fost respectate.
7. Identificaţi în chestionarul WVS 2012 variabile care pot fi folosite în procesul de validare logică. Verificaţi dacă există situaţii în care logica a fost încălcată.

5. Gestionarea variabilelor

Despre acest subiect am mai discutat. Am învăţat să modificăm o variabilă folosind **Transform > Recode into Same Variables**. De asemenea, am învăţat să creăm o variabilă nouă folosind o funcţie sau o formulă, prin intermediul **Transform > Compute**. În acest capitol dezvoltăm acest subiect. O mare parte din activitatea de analiză cantitativă a datelor este destinată pregătirii variabilelor.

Voicu, Rusu şi Comşa (2013) vor să explice solidaritatea românilor. Solidaritatea este, pentru aceştia, o atitudine faţă de alte persoane care denotă cooperare, interes, preocupare, sprijin etc. Solidaritatea este măsurată printr-un scor compozit obţinut prin cumularea răspunsurilor la mai mulţi itemi. Factorii care determină solidaritatea sunt orientarea postmaterialistă sau materialistă, religiozitatea şi comportamentul religios, identificarea naţională, încrederea generalizată, individualismul, clasa socială, vârsta, venitul, educaţia, sexul şi tipul de localitate de rezidenţă. Analiza prin care doresc să testeze ipotezele este regresia liniară multiplă, tehnică prezentată într-un capitol al acestei cărţi. Observăm că modelul explicativ propus de autori este destul de complex. Fiecare variabilă din model, începând cu cea dependentă (solidaritatea), trebuie pregătită pentru analiză. Pregătirea se va face ţinând cont şi de caracteristicile pe care le pot avea variabilele în analiza de regresie liniară. Din acest motiv, informaţiile despre cum gestionăm variabilele capătă un rol esenţial în procesul de analiză cantitativă.

În acest capitol vom discuta despre meniul **Transform > Recode into Different Variables** şi vom afla câteva informaţii noi despre meniul **Transform > Compute**.

5.1. Crearea unei alte variabile utilizând meniul Recode into Different Variable

Înainte de a începe analiza datelor, vă recomand să salvaţi într-un loc sigur baza de date în forma primită de la cei care au curăţat-o. Aceasta va fi baza de referinţă la care apelăm atunci când am pierdut informaţii din copia pe care lucrăm.

De exemplu, eu prefer să şterg codurile de nonrăspuns din baza de date, lăsând celulele respective goale. Astfel SPSS le va trata tot timpul ca **missing values** : nu mai există riscul să obţin rezultate greşite, pentru că am uitat să le definesc

în coloana **Missing** din **Variable View** sau folosind comanda **MISSING VALUES** din sintaxă. Această preferinţă poate fi satisfăcută folosind meniul **Transform > Recode into Same Variables**. Dar odată rulată comanda, am şters acea informaţie din variabilele cu care lucrez. Dacă, ulterior, doresc să realizez un profil al celor care au declarat că nu ştiu răspunsul la o întrebare şi să îl compar cu profilul celor care nu vor să răspundă la aceeaşi întrebare, atunci nu mai pot face acest lucru. Să ne gândim la toate veniturile unei persoane. Aceasta poate să câştige bani din salariul la principalul loc de muncă, dar poate avea şi un loc de muncă secundar, unde lucrează pe proiect. De asemenea, poate avea un cont de economii şi astfel primeşte lunar o dobândă. Toate aceste venituri se adună şi rezultă venitul lunar total al persoanei respective. Cercetătorul doreşte să estimeze venitul mediu al românilor pentru luna februarie a anului 2014. Va pune în chestionar o rubrică de forma : „Dacă adunaţi veniturile din toate sursele, vă rog să îmi spuneţi câţi bani aţi câştigat dvs. personal în luna februarie". Respondentul trebuie să aproximeze o sumă dacă nu o cunoaşte pe cea exactă. În teren, unii respondenţi ne oferă un răspuns. Alţii, în schimb, refuză să facă acest lucru. Motivele sunt multiple : lucrează „la negru", operatorul nu le inspiră încredere să declare o informaţie atât de personală etc. În fine, cei care sunt mai puţin preocupaţi de gestionarea lunară a veniturilor lor s-ar putea să nu ştie şi, decât să ofere un răspuns greşit, preferă să aleagă această variantă de răspuns. În baza de date, la introducere, fiecare dintre aceste situaţii primeşte codul corespunzător. Deci variabila va avea valori de la 0 la cel mai mare venit şi codurile 98 (nu ştiu) şi 99 (nu răspund). Ipoteza mea este că cei care au declarat că nu ştiu au alte caracteristici decât cei care au refuzat să răspundă. Dacă vreau să le compar caracteristicile şi am şters codurile, fără să fi păstrat o copie a bazei originale, nu mai pot face acest lucru. Concluzia : poate ar fi fost mai bine să creez o variantă nouă în care am şters codurile de nonrăspuns, păstrând-o şi pe cea iniţială. Să nu confundaţi ceea ce povestesc aici cu definirea nonrăspunsurilor din capitolul precedent. Definirea nonrăspunsurilor presupune că am păstrat codurile lor, numai că le dezactivăm din analizele pe care le rulăm. Eu vorbesc despre ştergerea fizică din bază.

O altă situaţie în care putem folosi **Recode into Different Variables** este atunci când vrem să prezentăm un tabel care conţine încrucişarea dintre o variabilă măsurată metric, cum ar fi vârsta, şi încrederea în oameni. Vârsta este înregistrată în ani împliniţi : 18, 19, 20 etc. Încrederea în oameni este înregistrată folosind două variante de răspuns : „se poate avea încredere în cei mai mulţi oameni" sau „e mai bine să fii atent în relaţiile cu oamenii". Dacă am realiza un tabel de contingenţă între cele două variabile, ar fi inutil, pentru că vârsta are foarte multe valori (în WVS 2012 pentru România, între 18 şi 85 de ani). În această situaţie, alegem să recodificăm vârsta, adică să creăm o variabilă cu câteva categorii alese după un criteriu teoretic întemeiat stabilit de cercetător.

Am putea alege categoriile : 18-29, 30-39, 40-49, 50-59, 60+. Observăm că pierdem informaţie. Aducem în aceeaşi categorie persoane de vârste diferite. De aceea categoriile nu se fac la întâmplare, ci motivat. Persoanele care sunt incluse în aceeaşi categorie trebuie să aibă trăsături comune, dar şi diferite faţă de ale persoanelor din celelalte categorii. Dacă folosim **Recode into Same Variables**, pierdem informaţia detaliată şi nu mai avem cum să o recuperăm. Dacă folosim **Recode into Different Variables**, creăm o variabilă nouă care conţine categoriile respective şi o păstrăm pe cea originală cu toată informaţia. Pe parcursul analizei, s-ar putea să ne trebuiască alte categorii de vârstă, să zicem 18-24, 25-29 etc. Putem să le obţinem oricând, pentru că avem variabila iniţială. Aici trebuie să facem trimitere către procesul de elaborare a întrebărilor din chestionar. Există cele patru niveluri de măsurare teoretice : nominal, ordinal, interval şi raport. Sintetic, proprietăţile acestora sunt prezentate în tabelul 5.1.

Tabelul 5.1. Niveluri de măsurare

	Nominal	**Ordinal**	**Interval**	**Raport**
Categorii	Da	Da	Da	Da
Categorii ordonate		Da	Da	Da
Distanţa dintre categoriile ordonate este egală			Da	Da
Număr				Da

Nivelul de măsurare cel mai slab din punct de vedere statistic este cel nominal. Sexul are două categorii : masculin şi feminin. Suntem obişnuiţi ca acestora să le fie atribuite codurile 1 şi 2. Dar codurile acestea puteau fi foarte bine înlocuite cu 1001 şi 23. Nu avea nici o importanţă. Sexul feminin nu este pe locul 2, pentru că primeşte codul 2, după nici un criteriu. La fel, sexul masculin nu este pe locul 1, pentru că primeşte codul 1, tot după nici un criteriu. O discuţie detaliată a acestor concepte este întâlnită în orice manual de metodologie cantitativă sau de statistică. Aş sublinia doar această idee : dacă puteţi măsura o variabilă folosind un nivel de măsurare de raport, atunci faceţi acest lucru. Dacă nu se poate utiliza un nivel de măsurare de raport, atunci căutaţi să folosiţi unul de interval sau măcar ordinal. Dacă nici acest lucru nu este posibil, atunci folosiţi un nivel nominal. Dintr-un număr putem face oricâte categorii şi de orice fel dorim. Din categorii nu putem face numere. Am văzut deseori chestionare în care vârsta este înregistrată sub formă de categorii. Oricât de detaliate ar fi, tot se pierde informaţie. Aş măsura o variabilă care este metrică sub formă de categorii, doar dacă mă ajută să reduc numărul de non-răspunsuri. Dar, în chestionar, aş utiliza ambele variante. De exemplu, aş întreba care este venitul din luna trecută, lăsând posibilitatea să declare o sumă şi apoi aş întreba şi în ce categorie se încadrează. Evident, operatorul de teren, dacă

a aflat suma, o va încadra singur în categoria aferentă. Dar dacă nu a aflat suma, datorită refuzului, poate aflǎ măcar categoria.

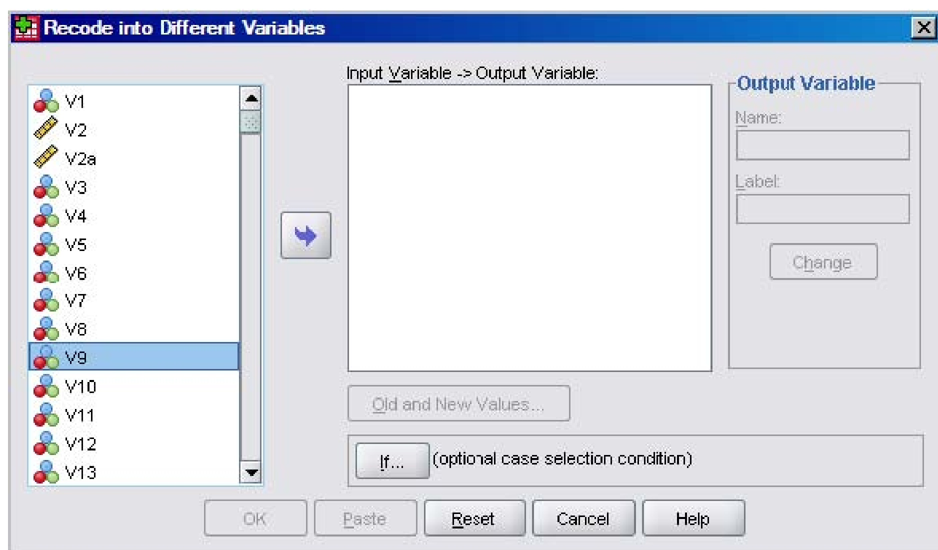
Variabilele dummy sunt un alt exemplu de utilizare a meniului **Recode into Different Variables**. O variabilă dummy ia valorile 1 sau 0. Valoarea 1 este atribuită caracteristicii care ne interesează, iar valoarea 0 celeilalte sau celorlalte. În regresia liniară muliplă nu pot folosi sexul codificat cu 1 şi 2. Aleg cine primeşte 1, să zicem bărbaţii, iar 2 va fi transformat în 0. Dacă vreau să folosesc starea civilă ca predictor al fericirii şi presupun că fenomenul explicat variază diferit pentru cei căsătoriţi şi pentru cei care au sau nu o relaţie, procedez astfel :

- presupun că cei căsătoriţi sunt cei mai fericiţi, aşadar voi alege drept referinţă această categorie. Pentru ea nu mai creez un dummy ;
- creez un dummy în care 1 este atribuit celor care au o relaţie, dar nu sunt căsătoriţi, iar 0 le este atribuit celor care nu au o relaţie, dar şi celor căsătoriţi ;
- creez un al doilea dummy în care 1 este atribuit celor care nu au o relaţie, fie sunt divorţaţi, fie sunt văduvi, iar 0 este atribuit celor care nu au o relaţie şi celor care sunt căsătoriţi.

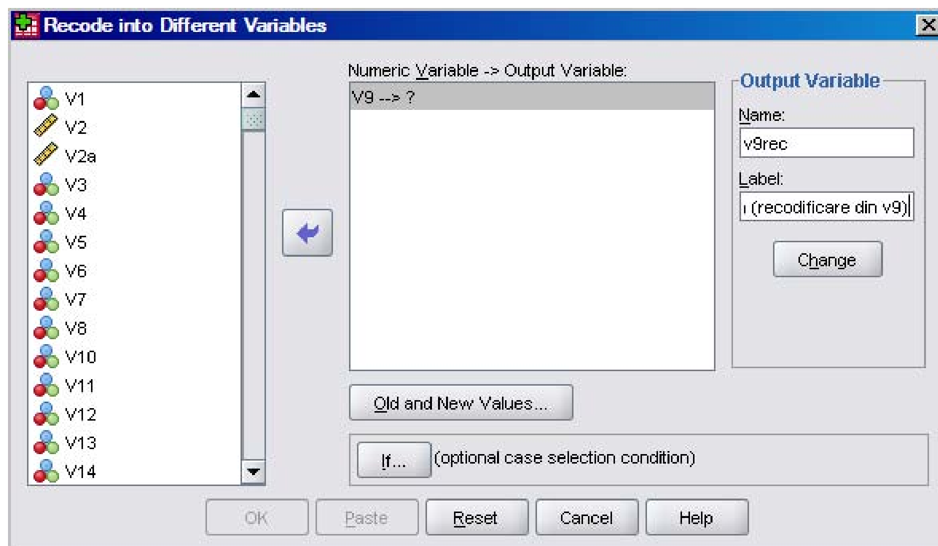
Voi prezenta meniul, folosind o altă situaţie care poate fi întâlnită în procesul de analiză : inversarea scalei de răspuns. În analiza elaborată de Voicu, Rusu şi Comşa (2013), religiozitatea este măsurată ca gradul de importanţă acordată religiei. În chestionarul folosit, întrebarea este „Vă rugăm să ne spuneţi cât de importante sunt următoarele lucruri în viaţa dumneavoastră : ...religia ? ” şi are patru variante de răspuns, de la „foarte importantă” (codul 1) la „deloc importantă” (codul 4). Pentru ca rezultatul analizei de regresie să fie mai uşor de citit, autorii au inversat scala astfel încât codul mare (4) să fie atribuit etichetei pozitive („foarte importantă”), iar codul mic (1) să fie atribuit etichetei negative („deloc importantă”). Măsura solidarităţii este orientată similar : un scor mare indică solidaritate ridicată. Astfel, dacă atunci când creşte religiozitatea creşte şi solidaritatea, coeficienţii de regresie vor avea semnul plus, iar interpretarea va fi intuitivă. Accesând meniul **Transform > Recode into Different Variables**, se deschide fereastra din figura 5.1a. Structura ferestrei ne este parţial familiară, pentru că seamănă cu cea de la **Recode into Same Variables**. Butonul **If** este folosit dacă dorim să punem o condiţie care să fie activă atunci când creăm variabila nouă. În secţiunea **Output Variable**, care iniţial este inactivă, introducem un nume pentru variabila pe care o creăm (**Name**) şi o etichetă care explică numele (**Label**). Completarea informaţiei la **Name** este obligatorie. La **Label** este opţională, dar recomandată. Altfel ar trebui să mergem în **Variable View** la coloana **Label** sau în sintaxă şi să folosim comanda **VARIABLE LABELS**. Figura 5.1b prezintă cum se modifică interfaţa.

Figura 5.1. Meniul Transform > Recode into Different Variables

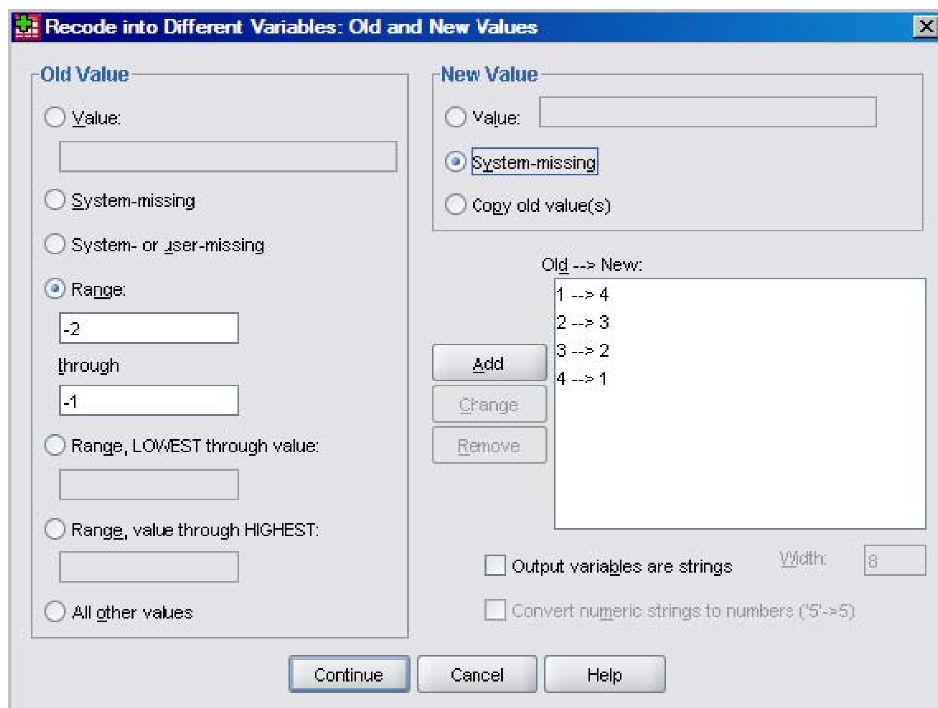
(a)



(b)



(c)



După ce introducem variabila V9 din lista de variabile din stânga în secţiunea **Numeric Variable --> Output Variable**, se activează **Name** şi **Label**. Numele trebuie să respecte regulile programului. Eu prefer să adaug în coada numelui variabilei iniţiale expresia „rec" de la „recodificată". Opţiunea dumneavoastră poate fi alta. La **Label** prefer să pun în etichetă informaţia „recodificare din variabila iniţială". Astfel, am o evidenţă clară a variabilelor pe care le-am creat. Odată completate aceste informaţii, trebuie să apăsăm butonul **Change**. Făcând acest lucru, dispare semnul de întrebare (**V9 --> ?**) şi apare : Numeric Variable -> Output Variable: V9 --> v9rec .

Următorul pas presupune să modificăm codurile conform nevoilor de analiză. Apăsăm butonul **Old and New Values** şi se deschide fereastra din figura 5.1c. Fereastra are trei secţiuni : **Old Value**, **New Value** şi cea care ne arată ce transformări facem. Mai întâi, trebuie să introducem în fereastra **Old Value** codurile variabilei iniţiale pe care le dorim transformate într-un fel sau altul. Aici dorim să inversăm scala : 1 devine 4, 2 devine 3, 3 devine 2, 4 devine 1. Se impune transformarea codurilor unul câte unul. O să lucrăm cu **Value**. După fiecare transformare, apăsăm butonul **Add**. Variabila V9 are şi două coduri de nonrăspuns, –2 şi –1. Pentru că nu vrem să le păstrăm în variabila nouă, le vom defini **System-missing**. Introducem la **Range** –2 la –1 şi bifăm **System-missing**.

Variabila independentă pentru autoevaluarea stării de sănătate, care are numele V11 în baza de date, are distribuţia din tabelul 5.2. O persoană nu a oferit un răspuns valid. Codurile sunt etichetate invers decât îmi doresc : codul mic (1) este asociat etichetei pozitive, iar codul mare (4) este asociat etichetei negative. Pentru că dorim să interpretăm efectul pozitiv al autoevaluării sănătăţii asupra satisfacţiei cu viaţa, aşteptându-ne la o relaţie pozitivă (semn + la coeficientul de regresie), recodificăm variabila astfel încât codurile să fie în acord cu etichetele. De asemenea, o să ştergem din bază nonrăspunsul respectiv.

Tabelul 5.2. Tabel de frecvenţă pentru autoevaluarea stării de sănătate

V11 State of health (subjective)					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very good	232	15.4	15.4	15.4
	2 Good	747	49.7	49.7	65.2
	3 Fair	390	25.9	25.9	91.1
	4 Poor	134	8.9	8.9	100.0
	Total	1502	99.9	100.0	
Missing	-2 No answer	1	.1		
Total		1503	100.0		

Aşadar, mergem în meniul **Transform > Recode into Different Variables**. Dacă nu am închis baza de date între timp, o să observăm comenzile de la recodificarea anterioară. Pentru a nu ne complica, apăsăm butonul **Reset**, acesta ştergând orice informaţie care era prezentă în meniul respectiv. Trecem V11 în dreapta. Îi dăm un nume : v11rec. Etichetăm numele : autoevaluarea stării de sănătate (recodificare din V11). Apăsăm **Change**. Apăsăm **Old and New Values**. Transformăm pe rând fiecare valoare folosind **Value** de la **Old Value : 1 --> 4, 2 --> 3, 3 --> 2, 4 --> 1, -2 --> System-missing**. Apăsăm butonul **Add** după fiecare modificare. **Continue** şi **OK**.

După recodificare trebuie să verificăm dacă am lucrat corect. În cazul creării unei noi variabile prin recodificare, vom realiza un tabel de contingenţă (**Crosstab**) dintre variabila iniţială (V9 sau V11) şi variabila nou-creată (v9rec sau v11rec). Acest tabel se realizează din meniul **Analyze > Descriptive Statistics > Crosstabs**. Pe rând (**Row**) introducem variabila cu mai multe categorii. Pe coloană (**Column**) introducem variabila creată (tabelul 5.3).

În primul rând, observăm că nu avem etichete pentru codurile variabilei nou-create : 1, 2, 3 şi 4. Deci trebuie să le introducem în coloana **Values** din **Variable View** sau folosind sintaxa de mai jos. Trebuie reţinut că acesta este, de cele mai multe ori, primul pas după recodificare.

VALUE LABELS v9rec
1 not at all important
2 not very important
3 rather important
4 very important

După ce am rulat această sintaxă, realizăm din nou tabelul. Rezultatul este vizibil în tabelul 5.3b. Acum este mai uşor de citit. Al doilea lucru pe care îl observăm este că nu mai apar coduri de nonrăspuns. Dacă lucrăm cu **user-missing** sau **system missing**, comanda **Crosstabs** le va ignora. Pe noi ne interesează, în acest tabel, să vedem dacă etichetelor le corespund oamenii potriviţi. Aceştia sunt distribuiţi pe diagonală, deci am lucrat corect. Atenţie însă : dacă am etichetat greşit, programul nu ne va avertiza. Să fim atenţi la fiecare etapă de lucru.

Tabelul 5.3. Tabel de contingenţă pentru verificarea corectitudinii recodificării

(a)

V9 Important in life: Religion * v9rec importanta religiei in viata (recodificare din v9) Crosstabulation

Count

		v9rec importanţa religiei în viaţă (recodificare din v9)				Total
		1	2	3	4	
V9 Important in life: Religion	1 Very important	0	0	0	758	758
	2 Rather important	0	0	500	0	500
	3 Not very important	0	192	0	0	192
	4 Not at all important	48	0	0	0	48
Total		48	192	500	758	1498

(b)

V9 Important in life: Religion * v9rec importanţa religiei în viaţă (recodificare din v9) Crosstabulation

Count

		v9rec importanta religiei in viata (recodificare din v9)				Total
		1 not at all important	2 not very important	3 rather important	4 very important	
V9 Important in life: Religion	1 Very important	0	0	0	758	758
	2 Rather important	0	0	500	0	500
	3 Not very important	0	192	0	0	192
	4 Not at all important	48	0	0	0	48
Total		48	192	500	758	1498

5.2. Crearea unei alte variabile utilizând meniul Compute

Am discutat deja o situaţie în care folosim **Transform > Compute**. Pentru că mi se pare important, am să insist prezentând, pentru început, cum realizăm o variabilă de ponderare (**weight**).

Să presupunem că ponderăm în funcţie de mediul de rezidenţă, vârstă şi sex. Mai întâi trebuie să stabilim care sunt categoriile pentru fiecare criteriu. Categoriile se aleg şi în funcţie de cum este disponibilă informaţia pentru acestea. Am ales categoriile urban şi rural pentru mediul de rezidenţă şi categoriile 18-34, 35-49, 50-64, 65+ pentru vârstă. Pentru sex avem doar două categorii : bărbat sau femeie. Căutăm la Institutul Naţional de Statistică informaţii pentru tabelul :

Vârstă	**Bărbaţi în Urban**	**Femei în Urban**	**Bărbaţi în Rural**	**Femei în Rural**	**Total**
18-34	…	…	…	…	**…**
35-49	…	…	…	…	**…**
50-64	…	…	…	…	**…**
65+	…	…	…	…	**…**
Total	…	…	…	…	**…**

Mai concret, informaţiile pe care trebuie să le punem în fiecare celulă sunt numerele de persoane care se încadrează simultan în toate cele trei categorii desemnate de rândurile şi coloanele tabelului. De exemplu, bărbaţii care locuiesc în urban şi au vârsta cuprinsă între 18-34 de ani ar putea fi în număr de 1.600.000. Realizăm acelaşi tabel şi pentru eşantion. Evident, numerele din fiecare celulă vor fi mult mai mici, dată fiind mărimea eşantionului. De exemplu, în eşantion ar putea fi incluşi 106 bărbaţi care locuiesc în urban şi au vârsta cuprinsă în intervalul 18-34 de ani

Calculăm proporţia fiecărei celule din totalul populaţiei, respectiv a eşantionului. Vor rezulta două noi tabele care conţin aceste proporţii. Apoi vom împărţi proporţiile din populaţie la proporţiile din eşantion :

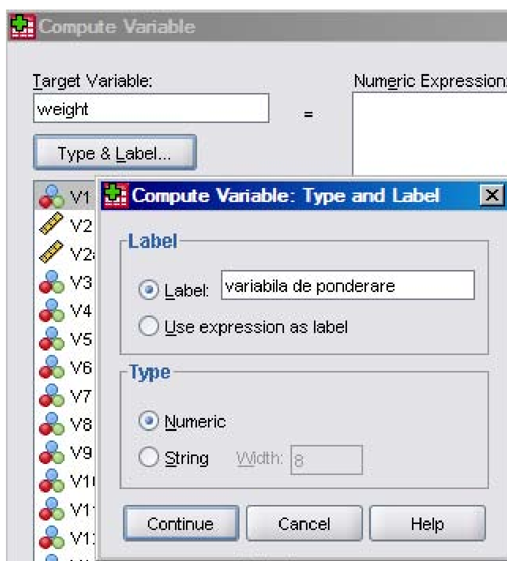
Vârstă	**Bărbaţi în Urban**	**Femei în Urban**	**Bărbaţi în Rural**	**Femei în Rural**	**Total**
18-34	9,16/7.07=1,2942	…			
35-49	…				
50-64					
65+					
Total					

Rezultatul final reprezintă valorile pe care le va lua ponderea pentru fiecare dintre aceste categorii compuse. Această nouă variabilă trebuie introdusă în SPSS. Realizăm acest lucru cu meniul **Transform > Compute**. În secţiunea **Target**

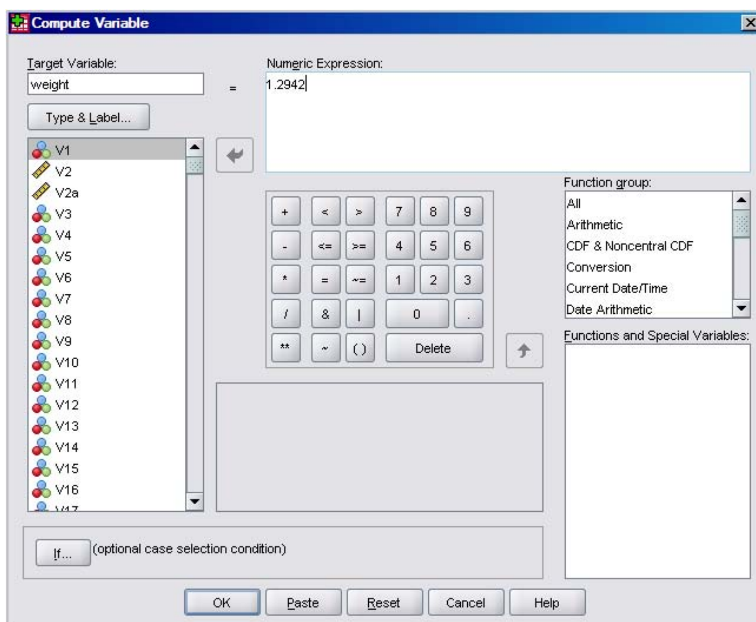
Variable, dăm un nume variabilei pe care o realizăm, căreia îi atribuim şi o etichetă în secţiunea **Label** din fereastra care se deschide apăsând butonul **Type & Label** (figura 5.2a).

Figura 5.2. Transform > Compute : crearea unei variabile de ponderare

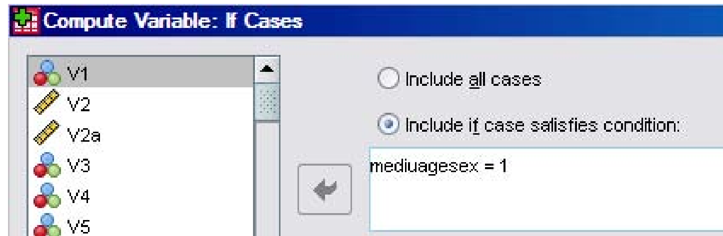
(a)



(b)



(c)

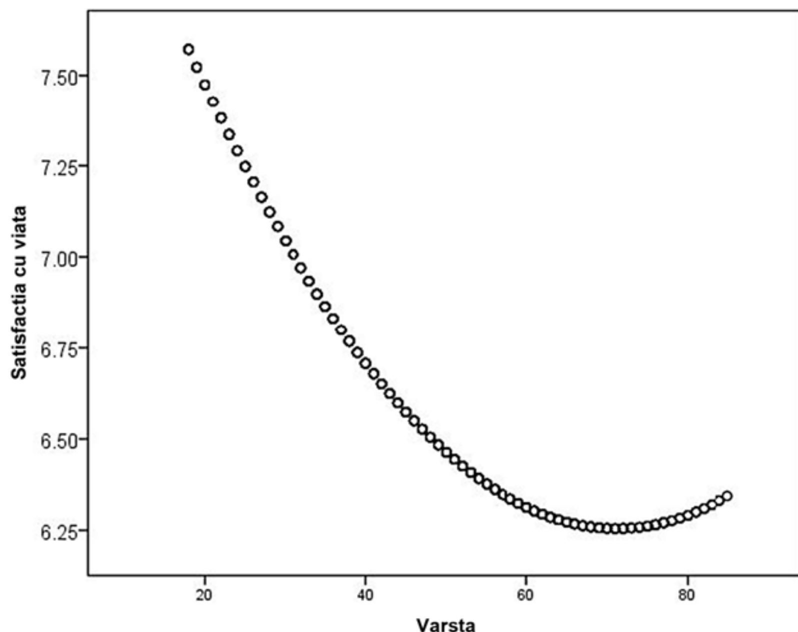


Observăm semnul „=" în dreapta câmpului **Target Variable**. În dreapta acestui semn, în câmpul **Numeric Expression**, introducem formula prin care realizăm noua variabilă. În cazul de faţă, nu avem o formulă: doar imputăm valoarea 1.2942 pe care o introducem fie din tastatură, fie folosind butoanele din centrul ferestrei (figura 5.2b). Dacă apăsăm **OK** acum, variabila de ponderare (**weight**) va avea valoarea 1.2942 pentru toate persoanele din eşantion. Însă această pondere este doar pentru categoria bărbaţilor care locuiesc în urban şi au vârsta în intervalul 18-34 de ani. De aceea trebuie să folosim şi butonul **If...** din colţul stânga jos al ferestrei. Apăsând acest buton se deschide fereastra din figura 5.2c.

Iniţial este bifată opţiunea **Include all cases**. Pentru că vrem să punem o condiţie, vom bifa **Include if case satisfies condition** şi vom introduce condiţia în câmpul activat. În exemplul nostru, am presupus că în baza de date există deja o variabilă care reflectă apartenenţa simultană la cele trei categorii. Aceasta are numele mediuagesex şi conţine 16 categorii. Codul 1 reprezintă categoria bărbaţi care locuiesc în mediul urban şi au vârsta între 18 şi 34 de ani. Pentru că ponderea 1.2942 este ponderea pentru această categorie, atunci vom introduce aici condiţia **mediuagesex = 1**. Astfel SPSS va atribui ponderea 1.2942 doar categoriei 1 de la variabila mediuagesex. Repetăm procedura pentru toate celelalte categorii.

Un alt exemplu. Relaţia dintre satisfacţia cu viaţa şi vârstă nu este liniară (Lelkes, 2008). Adică satisfacţia nu creşte/descreşte, constant, odată cu înaintarea în vârstă. Mai degrabă, cele două au o relaţie nonlineară asemănătoare cu cea reprezentată în figura 5.3. Cel mai înalt nivel al satisfacţiei cu viaţa este trăit în tinereţe, când grijile materiale şi sociale nu sunt atât de multe, părinţii încă îi întreţin pe copii etc. Urmează momente cum ar fi cel al intrării pe piaţa muncii, al formării propriei familii, al accentuării independenţei financiare etc. Copiii pleacă de acasă, grijile cu privire la siguranţa locului de muncă se accentuează etc. Vine vârsta pensionării, grijile legate de profesie se reduc, dar apar probleme de sănătate asociate vârstei, moartea partenerului de viaţă etc. Pe de altă parte, oamenii îşi pot urmări interesele personale mai mult decât înainte, cel puţin prin prisma timpului liber de care dispun. Toate acestea sunt explicaţii plauzibile pentru acest tip de relaţie dintre vârstă şi satisfacţia vieţii.

Figura 5.3. Relaţie nonliniară dintre vârstă şi satisfacţia cu viaţa

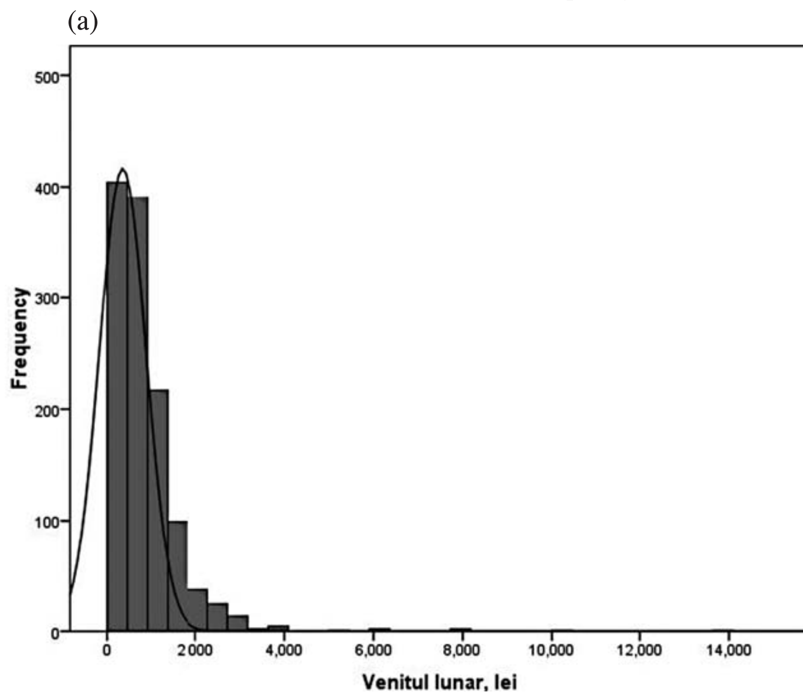


Dacă vrem să aplicăm o modelare liniară, trebuie să includem şi acel punct de inflexiune în analiză. Acest lucru se face aici prin introducerea variabilei vârstă ridicată la pătrat, alături de variabila vârstă iniţială. Variabila vârstă la pătrat va fi creată folosind meniul **Compute**. La **Numeric Expression** notăm formula : varsta2=V242 * V242. V242 este variabila care conţine vârsta respondentului din WVS 2012.

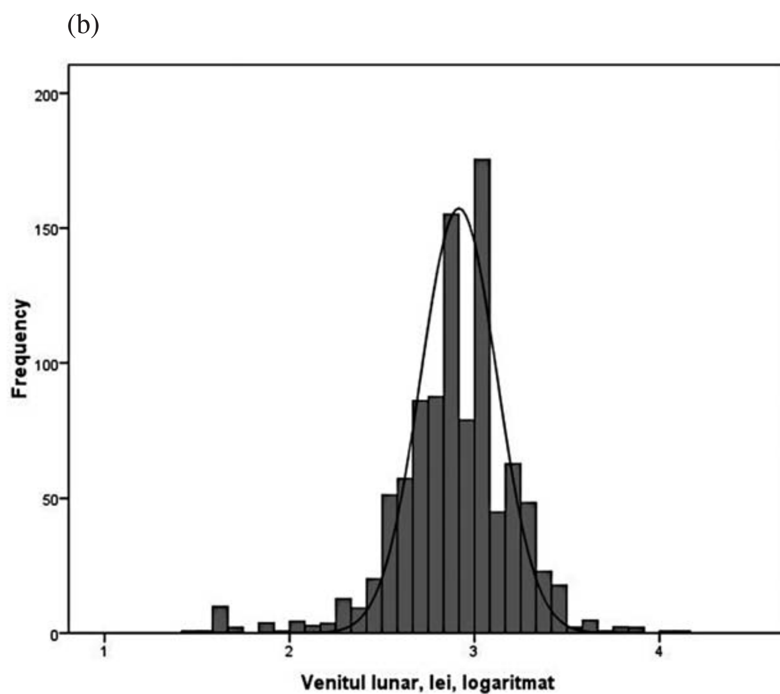
Astfel de transformări sunt frecvente în analizele multivariate. Una care foloseşte funcţiile implementate în SPSS presupune calcularea unui logaritm. Această transformare este frecvent întâlnită pentru variabila venit care nu are o distribuţie normală, ci, de regulă, alungită la dreapta (figura 5.4a). Majoritatea românilor au venituri mici, dar există şi români care au venituri ceva mai mari. Unii dintre aceştia pot să se îndepărteze destul de mult de majoritate. În analizele statistice, aceştia sunt consideraţi cazuri extreme (**outlieri**). Trebuie văzut în ce măsură afectează rezultatele analizelor statistice pe care le rulăm. Putem transforma variabila folosind una dintre funcţiile de logaritmare. La **Numeric Expression** aducem din secţiunea **Functions and Special Variables**, dând dublu click pe ea, funcţia **LG10()**. Trebuie doar să introducem între paranteze, în locul semnului de întrebare, variabila din baza de date care conţine informaţii despre venit : cs237a în WVS 2012. Funcţia devine **LG10(cs237a)**. Apăsăm **OK**. Distribuţia variabilei logaritmate aproximează mai bine forma aşteptată (figura 5.4b). Problema acestor transformări este creşterea gradului de dificultate a interpretării coeficienţilor de regresie atunci când, în locul unităţii de măsură a variabilei iniţiale, folosim logaritmi sau rezultatele altor funcţii matematice.

Figura 5.4. Distribuţia venitului înainte şi după logaritmare

(a)

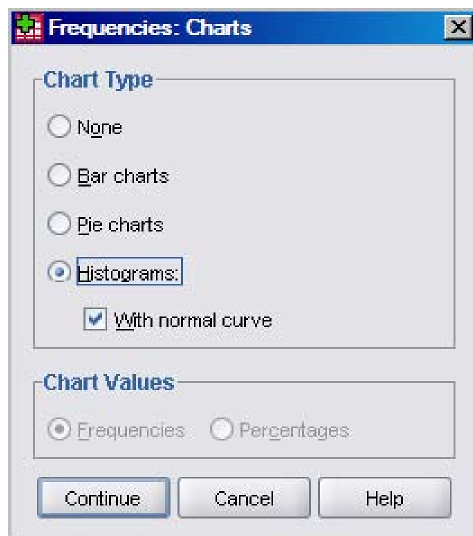


(b)



Histograma din figura 5.4 a fost creată din meniul **Analyze / Descriptive Statistics / Frequencies**. În fereastra care s-a deschis, apăsăm pe butonul **Charts** (figura 5.5). Iniţial este selectat **None**, dar noi suntem interesaţi de histogramă, de aceea facem selecţia corespunzătoare : **Histogram > With normal curve**.

Figura 5.5. Realizarea graficelor din meniul Frequencies > Charts



5.3. Exerciţii

Pentru aceste exerciţii utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultat(ă/e) în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a *Grupului Românesc pentru Studiul Valorilor Sociale* (http : //www.romanianvalues.ro).

1. Căutaţi pe siteul www.romanianvalues.ro newsletterul nr. 4 din 2013-2014 cu tema „Satisfacţia cu viaţa". Citiţi acest text şi realizaţi o listă cu variabilele utilizate în analiză.
2. Găsiţi variabila evaluarea stării de sănătate. Creaţi o variabilă dummy pornind de la aceasta. Căror coduri le atribuiţi valoarea 1 şi căror coduri le atribuiţi valoarea 0 ? Argumentaţi decizia.
3. Găsiţi variabila stare civilă. Creaţi o variabilă dummy pornind de la aceasta, astfel încât să reflecte categoriile „persoana are o relaţie" / „persoana nu are o relaţie".
4. Care sunt variabilele dummy pe care le puteţi crea din punct de vedere teoretic pornind de la variabila stare civilă ? Este fezabil să le creaţi pe toate ? Argumentaţi răspunsul.

5. Creaţi o variabilă care să conţină următoarele categorii de vârstă : 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65+.
6. Căutaţi în baza de date setul de variabile v102-v107 care se referă la încrederea în diferite categorii de persoane. Recodificaţi, creând variabile noi, toate aceste variabile astfel încât din patru variante de răspuns să rămâneţi doar cu două.
7. Citiţi lista de variabile v96-v101 din chestionar. Creaţi o nouă variabilă care să reprezinte suma tuturor acestor variabile. Ce măsoară această variabilă ? Cum interpretaţi scorul 60 ?
8. Căutaţi pe siteul www.romanianvalues.ro newsletterul nr. 2 din 2013-2014 cu tema „Încrederea în instituţii". Citiţi secţiunea „Cine are încredere în instituţiile politice ? ". Identificaţi în baza de date variabilele sex, vârstă, nivel de educaţie, mediu de rezidenţă, autopoziţionare în clasa socială şi mândria de a fi român. Recodificaţi aceste variabile astfel încât să corespundă modului în care sunt utilizate în această lucrare.

6. O primă privire asupra datelor

Prin cercetările noastre, căutăm să descriem şi/sau să explicăm un anumit fenomen social. Câţi români au emigrat în anul 2013 ? Dintre aceştia, câţi au studii superioare ? Câţi români suferă de o boală cronică ? Dintre aceştia, câţi au vârsta cuprinsă între 18 şi 30 de ani ? Câţi români consumă pufuleţi ? Dintre aceştia, câţi cumpără pufuleţii din hipermarket şi câţi din magazinul din proximitatea locuinţei ? În primul rând, descriem situaţia, dar, de regulă, vrem să şi explicăm de ce situaţia arată în felul acesta.

Testăm prezenţa unui efect (dacă..., atunci...), dar şi intensitatea cu care variază o variabilă dependentă în funcţie de variaţia variabilei independente (cu cât..., cu atât...). Putem compara nivelul de satisfacţie cu viaţa alromânilor care au emigrat cu cel al românilor care au decis să nu facă acest lucru. Dacă există diferenţe, emigrarea este factorul care produce diferenţa sau pot fi identificaţi şi alţi factori ? Nivelurile de satisfacţie cu viaţa sunt similare în cazul tuturor celor ce au emigrat sau variază în funcţie de caracteristicile ţării de destinaţie ? Care este factorul care creşte cel mai mult satisfacţia cu viaţa ? Sunt mai satisfăcuţi cu viaţa cei care au emigrat când erau mai tineri (sub 25 de ani) sau cei care au emigrat la o vârstă mai înaintată (peste 25 de ani) ?

Primul pas în acest demers este să ne familiarizăm cu datele. Să descriem modul în care gândeşte şi se comportă majoritatea. Primul pas este analiza statistică univariată. Avem o listă de variabile aleasă conform obiectivelor de cercetare şi, pentru fiecare dintre acestea, inspectăm distribuţiile şi diferiţi indicatori statistici care pot fi calculaţi pentru ele. Citim datele într-o manieră descriptivă. Dacă ne interesează să identificăm motivele pentru care unii români sunt mai fericiţi decât alţii sau motivele pentru care unii români îşi autoevaluează sănătatea ca fiind mai bună decât a altora, atunci începem prin a ne uita la distribuţia fericirii sau stării de sănătate a românilor la momentele alese pentru perspectiva cercetării. Câţi români sunt fericiţi şi câţi nefericiţi ? După care, trecem la analizele statistice bivariate. Începem să punem în relaţie variabilele din lista noastră, două câte două. Care sunt categoriile care cuprind cei mai mulţi români fericiţi : locuitorii oraşelor mici sau ai oraşelor mari, tinerii sau adulţii, cei căsătoriţi sau persoanele care nu au o relaţie de cuplu, cei care au absolvit facultatea sau cei care au absolvit doar liceul, cei din cuartila unu, doi sau trei de venit ş.a.m.d. ? Câţi români apreciază că starea lor de sănătate este bună şi câţi o apreciază ca fiind proastă ? Fenomenele sociale sunt complexe, de aceea analizele uni- sau bi-variate sunt insuficiente pentru a înţelege adecvat variaţia acestora. Orice analist doreşte

să ajungă la analizele statistice multivariate. Cine sunt cei care îşi evaluează sănătatea ca fiind mai bună : vegetarienii sau omnivorii, cei care merg la medic pentru controale preventive, cei mai educaţi, cei care fac sport ş.a.m.d. ?

Descrierea datelor se realizează prin calcularea unor indicatori statistici şi, vizual, prin inspectarea unor grafice. Calculăm indicatori ai tendinţei centrale, media (**mean**) şi mediana (**median**), dar şi indicatori ai variaţiei, abaterea standard (**standard deviation**) sau coeficientul de variaţie. Realizăm grafice bară (**bar chart**), histogramă (**histogram**) sau nor de puncte (**scatterplot**).

6.1. Cum gândeşte majoritatea şi cât de omogene sunt grupurile comparate

Indicatorii sintetici, cum sunt media sau mediana, oferă rapid, printr-un singur număr, o imagine de ansamblu asupra situaţiei majorităţii din populaţia de referinţă. Alţii, cum este abaterea standard, ne arată cât de omogene sunt, după aceeaşi caracteristică, diferite grupuri. Media şi mediana sunt indicatori ai tendinţei centrale. Abaterea standard este un indicator al variaţiei.

Aceşti indicatori pot fi calculaţi doar atunci când variabilele au anumite proprietăţi. Aceste proprietăţi sunt grupate sub numele de niveluri de măsurare (tabelul 5.1). Mediana este valoarea care împarte setul de date ordonate în două părţi egale. Poate fi calculată dacă variabila are cel puţin nivelul de măsurare ordinal sau, în limbajul cercetătorilor, este variabilă ordinală. Media poate fi calculată doar pentru variabile metrice, interval sau raport. Pentru variabilele nominale, vom inspecta distribuţia de frecvenţe : categoria cu cele mai multe unităţi va fi tendinţa centrală. Dacă ne reamintim coloanele din **Variable View**, mai exact coloana **Measure**, remarcăm că SPSS distinge între variabilele nominale (Nominal), ordinale (Ordinal) şi metrice (Scale). În cercetarea socială, atunci când aplicăm un chestionar, este destul de greu să măsori prin procedeele uzuale, la nivel de raport. În cel mai fericit caz, am reuşit să elaborăm variabile ordinale sau de interval. De aceea, în practică, pentru interval şi raport sunt folosite aproximativ aceleaşi analize statistice. O discuţie care clarifică multe dintre aceste aspecte este oferită de Agresti şi Finlay (2008).

În ştiinţele sociale, folosim frecvent media aritmetică pentru a reprezenta tendinţa centrală. Este larg cunoscută, majoritatea ştiind să o interpreteze. Spre deosebire de mediană, utilizează informaţia numerică din variabilă, nu doar ordinea scorurilor (Agresti şi Finlay, 2008). Totuşi, mediana este frecvent consultată de analist : este, cel puţin, un mecanism de verificare a mediei sau chiar înlocuitor al acesteia, atunci când datele conţin cazuri extreme (**outliers**). Cazurile extreme sunt persoane care au valori mult mai mari sau mult mai mici decât majoritatea la variabila respectivă. O persoană care are un salariu lunar de 25.000

de lei, în condiţiile în care următorul salariu, în ordine descendentă, este de 8.000 de lei, este un caz extrem. Aceasta nu este o situaţie ireală. Ea are însă un impact negativ asupra calculelor statistice. Prezenţa printre valorile variabilei chiar şi a unui singur caz extrem, indiferent că se află în partea stângă (valoare foarte mică) sau în partea dreaptă a scalei (valoare foarte mare), va afecta serios media, micşorându-i sau crescându-i foarte mult valoarea. Calculând salariul mediu folosind şi valoarea 25.000 lei va distorsiona media : salariul mediu va lua o valoare care nu reflectă situaţia majorităţii. Rotariu, Bădescu şi colaboratorii (2006), prezentând detaliat proprietăţile mediei şi medianei, atrag atenţia că media nu este valoarea mijlocie a seriei. Media se va încadra în intervalul valorilor variabilei pentru care este calculată, fiind exprimată în aceeaşi unitate de măsură cu aceasta. Dacă variabila este „salariu exprimat în lei”, atunci media va fi exprimată în lei. De Vaus (2002), la rândul său, subliniază un alt neajuns al mediei, care trebuie avut în considerare în momentul interpretării valorii calculate de program : aceeaşi medie poate fi obţinută din distribuţii diferite. Agresti şi Finlay (2008) demonstrează cum media este deplasată în direcţia cozii mai lungi, atunci când distribuţia este alungită la stânga sau la dreapta. Când grupurile pentru care este calculată sunt omogene, adică persoanele seamănă între ele, media va fi un indicator bun al tendinţei centrale, dar mai puţin bun atunci când grupurile sunt eterogene. Acesta este unul dintre motivele pentru care calculăm şi indicatori ai variaţiei sau dispersiei, împreună cu indicatorii tendinţei centrale.

Indicatorii variaţiei sau dispersiei arată gradul de împrăştiere sau omogenitate/eterogenitate a grupurilor investigate după o variabilă anume. Înainte de a calcula un indicator al variaţiei, trebuie să stabilim ce nivel de măsurare are variabila respectivă. Cel mai utilizat indicator este abaterea standard, care, pentru că foloseşte media în formula de calcul, poate fi calculat doar pentru variabile metrice. Putem compara abaterile standard calculate pentru aceeaşi variabilă în cazul a două grupuri. Grupul care arată cea mai mare abatere standard va fi mai eterogen. Dar această comparaţie nu ne va spune prea multe despre cât de omogen sau eterogen este fiecare grup. Agresti şi Finlay (2008) prezintă o regulă empirică aplicabilă distribuţiilor aproximativ normale, pe care o putem utiliza pentru a interpreta abaterea standard şi în termenii mărimii valorii acesteia : (1) aproximativ 68% dintre cazuri se află în intervalul [medie – abatere standard, medie + abatere standard], (2) aproximativ 95% dintre cazuri se află în intervalul [medie – 2 x abatere standard, medie + 2 x abatere standard] şi (3) aproape toate cazurile se află în intervalul [medie – 3 x abatere standard, medie + 3 x abatere standard]. Abaterea standard are câteva neajunsuri care pot fi corectate prin utilizarea altui indicator al variaţiei, coeficientul de variaţie. Coeficientul de variaţie este egal cu raportul dintre abaterea standard şi media variabilei. Acesta este util atunci când vrem să comparăm anumite grupuri (1) folosind o variabilă care are unităţi de măsură diferite şi/sau (2) nivelul general al valorilor variabilei este diferit în grupurile respective. Rotariu, Bădescu şi colaboratorii (2006) oferă o explicaţie detaliată în acest sens : nu poţi compara salariile din România, exprimate în lei, cu cele din Germania, exprimate în euro, la fel cum nu poţi compara masa corporală

a unor albine cu cea a unor elefanţi. Aceşti autori atrag atenţia la utilizările fără logică teoretică ale coeficientului de variaţie : poate fi calculat doar pentru nivelul de măsurare de raport, pentru că valorile au originea zero. De asemenea, nu trebuie utilizat pentru a compara grupurile folosind variabile care au conţinut diferit.

Sunt situaţii în care dorim să ştim ce procent din observaţii se află sub sau deasupra unei valori. Acest gen de informaţie ne este oferit, de exemplu, de mediană : 50% dintre observaţii se află sub această valoare şi 50% peste această valoare. Pentru informaţii mai detaliate utilizăm percentilele, întâlnite în cărţile de statistică sub denumirea de măsuri ale poziţionării (Agresti şi Franklin, 2013). Percentilele sunt de mai multe feluri. Cuartilele sunt foarte utilizate. Există trei cuartile, cuartila 2 fiind chiar mediana. Sub prima cuartilă se află 25% dintre cazuri, iar deasupra celei de-a treia cuartile se află tot 25% dintre cazuri. Cel mai simplu este să vă reprezentaţi o linie împărţită în patru segmente, fiecare segment reprezentând 25% din date. Asociată cuartilelor este abaterea intercuartilă, care ne arată distanţa dintre cuartilele trei şi unu. Din acest motiv, abaterea intercuartilă nu este influenţată de cazurile extreme, fiind utilizată pentru detectarea acestora : dacă o observaţie se află dincolo de 1.5 x AIQ, adică sub prima cuartilă sau peste a treia cuartilă, atunci s-ar putea să fie un caz extrem. Graficul **box-plot** ne ajută să vizualizăm acest gen de informaţii.

SPSS oferă mai multe posibilităţi prin care putem calcula indicatorii tendinţei centrale, variaţiei şi poziţionării.

Pentru variabilele nominale, utilizăm distribuţia de frecvenţe pe care o obţinem din meniul **Analyze > Descriptive statistics > Frequencies**. La întrebarea „În general vorbind, aţi spune că se poate avea încredere în cei mai mulţi oameni sau că e mai bine să fii atent în relaţiile cu oamenii ? ” adresată în WVS 2012 şi românilor, distribuţia răspunsurilor este cea prezentată în tabelul 6.1. În primul rând, remarcăm cele 15 persoane care nu au oferit un răspuns valid (coloana **Frequency**) (tabelul 6.1a). Trebuie să instruim programul că –2 şi –1 sunt coduri de nonrăspuns care trebuie dezactivate din analiză. Facem acest lucru fie în coloana **Missing** din **Variable View** (**Discrete missing values = -2, respectiv -1**), fie rulând sintaxa **MISSING VALUES V24 (-2, -1)**. Rezultatul este prezentat în tabelul 6.1b.

Tabelul 6.1. Tabel de frecvenţă : înainte şi după definirea nonrăspunsurilor

(a)

V24 Most people can be trusted					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-2 No answer	4	.2	.2	.2
	-1 Don´t know	11	.7	.7	1.0
	1 Most people can be trusted	115	7.7	7.7	8.6
	2 Need to be very careful	1373	91.4	91.4	100.0
	Total	1503	100.0	100.0	

(b)

V24 Most people can be trusted					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Most people can be trusted	115	7.7	7.7	7.7
	2 Need to be very careful	1373	91.4	92.3	100.0
	Total	1489	99.0	100.0	
Missing	-2 No answer	4	.2		
	-1 Don´t know	11	.7		
	Total	14	1.0		
Total		1503	100.0		

Observăm că 92% dintre români considerau, în 2012, că e mai bine să fii atent în relaţiile cu oamenii.

În acelaşi an, majoritatea românilor considerau că principala problemă din lume este sărăcia : 53% au ales această variantă de răspuns în defavoarea celorlalte (tabelul 6.2).

Tabelul 6.2. Tabel de frecvenţă : după definirea nonrăspunsurilor

V80 Most serious problem of the world					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 People living in poverty and need	782	52.0	52.7	52.7
	2 Discrimination against girls and women	105	7.0	7.1	59.8
	3 Poor sanitation and infectious diseases	205	13.7	13.9	73.7
	4 Inadequate education	260	17.3	17.5	91.2
	5 Environmental pollution	130	8.7	8.8	100.0
	Total	1483	98.7	100.0	
Missing	-2 No answer	7	.5		
	-1 Don´t know	12	.8		
	Total	20	1.3		
Total		1503	100.0		

Remarcaţi diferenţa dintre coloana **Percent** şi **Valid Percent**. În prima sunt calculate procentele luând ca bază întregul eşantion, adică şi pe cei care nu au

oferit un răspuns valid. În cea de-a doua sunt calculate procentele luând ca bază eşantionul valid, adică doar pe cei care au oferit un răspuns valid.

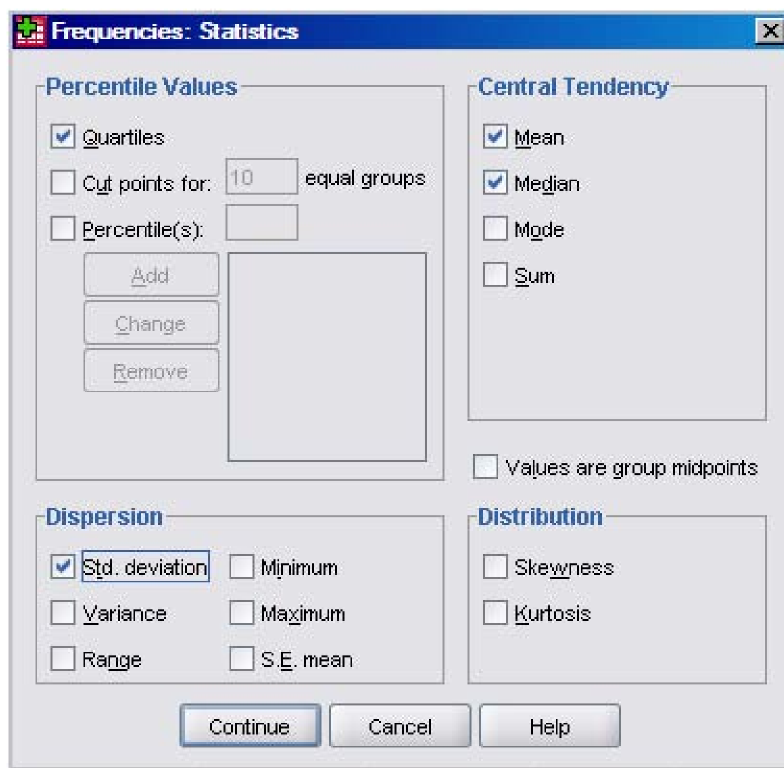
Pentru indicatorii tendinţei centrale, variaţiei şi poziţionării, calculabili pentru variabilele care au cel puţin nivelul de măsurare ordinal, putem utiliza meniurile **Analyze > Descriptive Statistics > Frequencies**, **Analyze > Descriptive Statistics > Descriptives** sau **Analyze > Descriptive Statistics > Explore**.

Meniul **Analyze > Descriptive Statistics > Frequencies** ne este deja familiar pentru că l-am folosit pentru a realiza tabelele de frecvenţă. Până acum, doar am introdus variabilele în partea dreaptă şi am apăsat butonul **OK**. Când utilizăm meniul, în fereastra care se deschide, observăm mai multe butoane. Cel care ne interesează aici este butonul **Statistics** (figura 6.1).

Acest submeniu ne permite să calculăm media, mediana, abaterea standard şi diferite tipuri de percentile. Pe lângă acestea, putem alege să calculăm şi alţi indicatori ai tendinţei centrale şi variaţiei cum ar fi modul, respectiv amplitudinea. De asemenea, în secţiunea **Distribution** putem calcula doi indicatori ai formei distribuţiei, **skewness** (alungirea) şi **kurtosis** (aplatizarea), dar despre aceştia discutăm la secţiunea de explorare a datelor.

Figura 6.1. Meniul Frequencies, butonul Statistics

(a)



(b)

Fereastra afişată prin apăsarea butonului **Statistics** este intuitivă. Observăm că indicatorii sunt grupaţi în secţiunile **Percentile Values** (poziţionare), **Central Tendency** (tendinţă centrală), **Dispersion** (variaţie) şi **Distribution** (forma distribuţiei). În analiza noastră, suntem interesaţi să cunoaştem tendinţa centrală pentru fericire şi sănătatea autoevaluată în rândul românilor. În baza de date WVS 2012, variabilele sunt V10 şi V11. Indicatorii statistici sunt prezentaţi în tabelul 6.3a, iar tabelele de frecvenţe sunt prezentate în tabelul 6.3b. Variabilele sunt ordinale : fericirea variază de la „deloc fericit” la „foarte fericit”, iar sănătatea autoevaluată variază de la „proastă” la „foarte bună”. Puteţi schimba ordinea în care sunt aşezate categoriile în funcţie de codurile lor dacă, în meniul **Frequencies**, apăsaţi butonul **Format** şi, în secţiunea **Order by**, bifaţi **Descending values**. Fiind variabile ordinale, putem calcula mediana şi măsurile poziţionării. În practică, deseori, întâlnim în multe lucrări şi medii calculate pentru acest tip de variabilă ordinală.

Tabelul 6.3. Tabele de frecvenţă şi indicatori statistici ai tendinţei centrale şi ai variaţiei

(a)

Statistics			
		V10 Feeling of happiness	V11 State of health (subjective)
N	Valid	1495	1502
	Missing	8	1
Mean		2.21	2.28
Median		2.00	2.00
Std. Deviation		.721	.830
Percentiles	25	2.00	2.00
	50	2.00	2.00
	75	3.00	3.00

(b)

V10 Feeling of happiness					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very happy	205	13.6	13.7	13.7
	2 Rather happy	833	55.4	55.7	69.4
	3 Not very happy	397	26.4	26.6	96.0
	4 Not at all happy	60	4.0	4.0	100.0
	Total	1495	99.5	100.0	
Missing	-2 No answer	4	.3		
	-1 Don´t know	4	.3		
	Total	8	.5		
Total		1503	100.0		

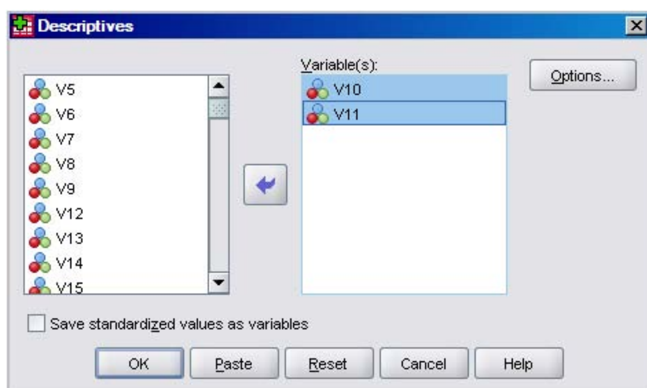
V11 State of health (subjective)					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very good	232	15.4	15.4	15.4
	2 Good	747	49.7	49.7	65.2
	3 Fair	390	25.9	25.9	91.1
	4 Poor	134	8.9	8.9	100.0
	Total	1502	99.9	100.0	
Missing	-2 No answer	1	.1		
Total		1503	100.0		

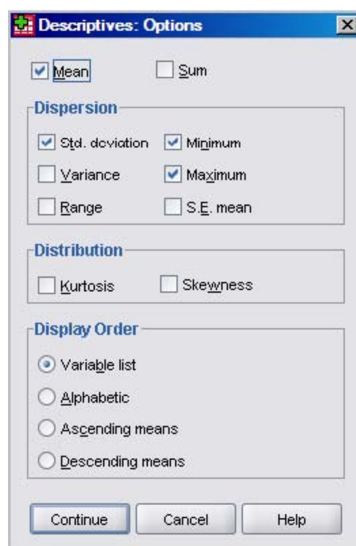
Mediana fericirii este egală cu 2, „destul de fericit”. Procentele ne arată că cel mai frecvent nivel de fericire ales de către români este „destul de fericit” (56%). Mediana stării de sănătate autoevaluate este egală cu 2, „bună”. Procentele ne arată că cea mai frecvent aleasă stare a sănătăţii de către români este „bună” (50%). Media este apropiată ca valoare de mediană pentru ambele variabile. Cuartilele

ne arată că cel puţin 25% dintre români au declarat că sunt „nu prea fericiţi” sau „deloc fericiţi” (percentila 75 = cuartila 3 = codul 3 „nu prea fericit”), respectiv că au o stare de sănătate „nu prea bună” sau chiar „proastă” (percentila 75 = cuartila 3 = codul 3 „nu prea bună”). Dacă dorim o informaţie mai detaliată, putem înlocui cuartilele cu decile, de exemplu : în secţiunea **Percentile Values** introducem valorile 10, 20,… , 100 (figura 6.1b). Agresti şi Finlay (2008) ne îndeamnă să fim precauţi cu interpretarea atunci când variabila are puţin categorii (variante de răspuns).

Un alt meniu din care putem obţine aceşti indicatori statistici este **Analyze > Descriptive Statistics > Descriptives**. Acesta este însă ceva mai limitat, permiţând doar calcularea mediei şi abaterii standard, fără mediană şi percentile. Odată intraţi în meniu (figura 6.2), apăsăm butonul **Options** şi alegem ce indicatori ne interesează. Rezultatele vor fi aceleaşi.

Figura 6.2. Meniul Descriptives

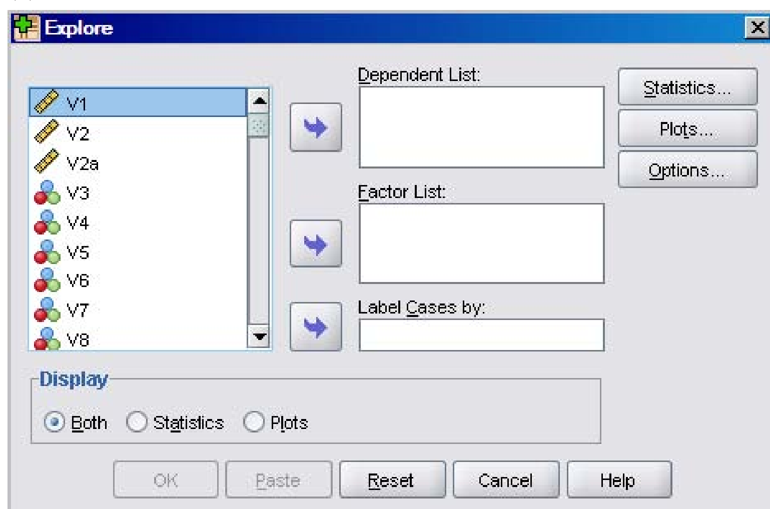




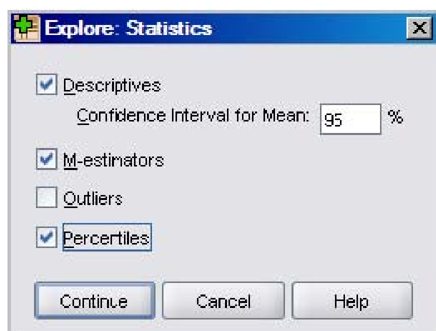
În fine, ultimul meniu prezentat aici, care poate fi folosit pentru calcularea acestor indicatori, este **Analyze > Descriptive Statistics > Explore** (figura 6.3a).

Figura 6.3. Meniul Explore

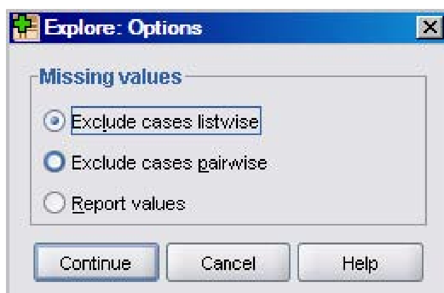
(a)



(b)



(c)



Acesta este ceva mai complex pentru că, aşa cum ne arată şi numele, este dedicat explorării datelor în vederea testării unor asumpţii de bază ale analizelor statistice uzual angajate în studiile sociale. Aici vom discuta doar despre cum obţinem indicatorii discutaţi, restul meniului fiind abordat în secţiunea dedicată explorării datelor.

Fereastra are două secţiuni care ne interesează în acest moment : **Dependent List** şi **Factor List**. La **Dependent List** introducem variabila pentru care dorim să calculăm statisticile, de exemplu, fericirea sau starea de sănătate autoevaluată. La **Factor List** introducem variabila care conţine grupurile care urmează să fie comparate, de exemplu, mediul de rezidenţă, care conţine două grupuri : locuitorii din urban şi locuitorii din rural. Apăsând butonul **Statistics** putem selecta, pe lângă statisticile descriptive discutate, **M-estimators**, **Outliers** şi **Percentiles** (figura 6.3b).

M-estimators sunt alternative robuste la medie şi mediană. Bifând **Percentiles** obţinem percentilele 5, 10, 25, 50, 75, 90 şi 95. Prefer să lucrez cu meniul **Frequencies**, pentru că îmi dă mai multă libertate în opţiuni. Bifând **Outliers**, ne oferă un tabel cu ceea ce SPSS consideră a fi caz extrem (tabelul 6.4). Acest tabel nu este foarte informativ, pentru că oferă doar o selecţie a aşa-ziselor valori extreme. Coloana **Case Number** conţine numărul rândului din **Data View**. Dacă am fi introdus în secţiunea **Label cases by** din fereastra principală (vezi figura 6.3a) o variabilă care conţinea id-ul unic al fiecărui respondent, atunci tabelul ar mai fi conţinut o coloană cu numele variabilei respective. Această alternativă este mai bună pentru că, dacă decidem să sortăm baza de date altfel decât în momentul în care am realizat tabelul (meniul **Sort Cases**), atunci informaţia din tabel devine inutilă.

Tabelul 6.4. Tabel Outliers obţinut din meniul Explore

Extreme Values				
			Case Number	Value
V10 Feeling of happiness	Highest	1	621	4
		2	622	4
		3	623	4
		4	624	4
		5	625	4[a]
	Lowest	1	769	1
		2	768	1
		3	767	1
		4	766	1
		5	765	1[b]
a. Only a partial list of cases with the value 4 are shown in the table of upper extremes.				
b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.				

În cadrul meniului **Explore**, o altă comandă care ne interesează acum este cea declanşată de butonul **Options** (figura 6.3c). Aici decidem cum sunt tratate nonrăspunsurile atunci când introducem, simultan, cel puţin două variabile la

Dependent List sau cel puţin două variabile la **Factor List**. Implicit, SPSS va trata nonrăspunsurile **listwise**, adică va dezactiva în analiză cazurile care au nonrăspunsuri. Decizia aparţine însă cercetătorului.

Tabelul cu statistici oferit de meniul **Explore** conţine multe informaţii utile (tabelul 6.5). Am calculat media, intervalul de încredere în jurul mediei, media calculată excluzând extremele distribuţiei (5% **Trimmed Mean**), mediana, varianţa (pătratul abaterii standard), abaterea standard, valoarea minimă pe care o ia variabila, dar şi valoarea maximă, amplitudinea (**range**), abaterea intercuartilă, alungirea (**skewness**) şi aplatizarea (**kurtosis**).

Tabelul 6.5. Output produs de meniul Explore

Descriptives				
			Statistic	Std. Error
V10 Feeling of happiness	Mean		2.21	.019
	95% Confidence Interval for Mean	Lower Bound	2.17	
		Upper Bound	2.25	
	5% Trimmed Mean		2.19	
	Median		2.00	
	Variance		.519	
	Std. Deviation		.721	
	Minimum		1	
	Maximum		4	
	Range		3	
	Interquartile Range		1	
	Skewness		.305	.063
	Kurtosis		.006	.127
V11 State of health (subjective)	Mean		2.28	.021
	95% Confidence Interval for Mean	Lower Bound	2.24	
		Upper Bound	2.32	
	5% Trimmed Mean		2.26	
	Median		2.00	
	Variance		.688	
	Std. Deviation		.830	
	Minimum		1	
	Maximum		4	
	Range		3	
	Interquartile Range		1	
	Skewness		.365	.063
	Kurtosis		-.336	.127

Închei prin a atrage încă o dată atenţia asupra stabilirii corecte a nivelului de măsurare al variabilei pentru care calculăm indicatorii statistici. Acest lucru se face înainte de realizarea calculelor respective. Deşi aici am calculat media şi

abaterea standard pentru variabile ordinale de tip Likert cu patru categorii, acest lucru nu înseamnă că acceptăm cu uşurinţă rezultatul primit. Vom întâlni în multe lucrări publicate astfel de analize. Trebuie să fim critici şi să ne gândim cât de bine respectă cerinţele de calcul astfel de măsurători şi cât de interpretabil este rezultatul analizei.

6.2. Asocierea dintre variabile categoriale. Tabelul de contingenţă (Crosstabs)

După inspectarea individuală a variabilelor, vrem să vedem cum sunt asociate diferite variabile. De regulă, avem o variabilă a cărei variaţie dorim să o explicăm şi mai mai multe variabile despre care credem că o influenţează. Aici gândim bivariat. Cei din cuartila 1 de venit sunt mai mulţumiţi cu viaţa lor decât cei din cuartila 2 ? Cei care au absolvit liceul sunt mai mulţumiţi cu viaţa lor decât cei care au absolvit facultatea ? Intuim deja de ce este util să învăţăm şi tehnici de analiză multivariată. Venitul mai mare creşte posibilitatea de a satisface mai multe nevoi şi aspiraţii, cum ar fi nevoia pentru o locuinţă cu mai multe camere, pentru o maşină mai încăpătoare, pentru vacanţe mai lungi etc. Cei care au absolvit niveluri formale de învăţământ mai înalte au mai multe cunoştinţe, lucru care le permite să fie mai flexibili pe piaţa muncii, să gestioneze riscurile mai uşor, să fie mai permeabili la schimbare etc. Însă, până la construirea unui model multivariat, ne putem face o idee despre obiectul studiului nostru folosind analizele bivariate. Decizia de a cumpăra un brand de cafea depinde de loialitatea faţă de brand ? Dacă investigăm doar consumatori de cafea care nu sunt loiali nici unui brand, atunci când sunt la raftul de cafea, este culoarea ambalajului un factor de decizie pentru cumpărare ?

Relaţia dintre două variabile categoriale poate fi observată folosind tabelul de contingenţă (**Crosstabs**). Variabilele categoriale sunt nominale sau ordinale. Esenţial este ca, atunci când realizăm un tabel de contingenţă, ambele variabile să aibă puţine categorii, pentru ca în fiecare celulă a tabelului să avem un număr rezonabil de cazuri. Un tabel cu 20 de rânduri şi 10 coloane nu este util, pentru că, probabil, multe celule nu vor avea cazuri. Nu există o regulă care să specifice care este numărul optim de rânduri şi coloane.

Persoanele care au încredere în semenii lor sunt mai fericite decât persoanele care nu au încredere în aceştia ? Alţi cercetători pot să pună întrebarea în sens invers : persoanele care sunt mai fericite au mai multă încredere în semenii lor decât persoanele care sunt mai puţin fericite ? Sensul relaţiei este stabilit printr-o atentă documentare teoretică. Programul de statistică nu alege variabila dependentă. El doar oferă calculele şi graficele pe care le solicităm. Alegerea sensului relaţiei este un act teoretic realizat înainte de a trece efectiv la analizarea datelor

în program. Domeniul fericirii este un exemplu foarte bun în ceea ce priveşte ambiguitatea direcţiei : de la fericire la altceva sau de la altceva la fericire. Pentru sociologi este specifică mai degrabă a doua variantă : presupunem că fericirea este starea la care trebuie să ajungem, trebuind să identificăm factorii care ne ajută în acest sens.

Să vedem care este relaţia dintre încredere şi fericire. În WVS 2012, fericirea este măsurată prin întrebarea : „V10. Luând în considerare toate aspectele vieţii dvs., aţi spune că sunteţi : 1. Foarte fericit ; 2. Destul de fericit ; 3. Nu prea fericit ; 4. Deloc fericit ? ". Încrederea este măsurată prin întrebarea : „V24. În general vorbind, aţi spune că se poate avea încredere în cei mai mulţi oameni sau că e mai bine să fii atent în relaţiile cu oamenii : 1. Se poate avea încredere în cei mai mulţi oameni ; 2. E mai bine să fii atent în relaţiile cu oamenii ? ". Ne aşteptăm că persoanele care au încredere în majoritatea oamenilor, adică aleg varianta 1 la V24, să fie mai fericite, adică aleg variantele 1 sau 2 la V10.

Avem două variabile categoriale : una nominală, V24, pe care o considerăm independentă, şi una ordinală, V10, pe care o considerăm dependentă. Adică V10 este influenţată de V24. Putem încrucişa aceste două variabile, pentru a vedea dacă presupunerea este corectă. Mai întâi realizăm câte un tabel de frecvenţă pentru V10 şi V24, pentru (1) a vedea dacă există coduri de nonrăspuns care nu sunt declarate **missing** în program şi pentru (2) a inspecta distribuţia variabilelor. Dacă există coduri de nonrăspuns nedeclarate **missing**, atunci trebuie să mergem în **Variable View** > coloana **Missing** şi să le declarăm. În ceea ce priveşte distribuţia, ne interesează să avem suficiente cazuri pentru fiecare variantă de răspuns de la cele două variabile. S-ar putea ca la fericire, V10, să fie necesară o recodificare care presupune gruparea categoriilor. Există oameni care nu experimentează nici un pic de fericire (aleg varianta 4 la V10) ? În această situaţie s-ar putea să dorim unirea categoriilor „deloc fericit" şi „nu prea fericit". De asemenea, s-ar putea ca la încredere, V24, să nu avem variaţie, adică majoritatea să aibă sau să nu aibă încredere în semenii lor. În această situaţie, avem mai multe posibilităţi : considerăm că încrederea nu este măsurată bine şi căutăm alt indicator pe care să îl folosim în analiză, sau considerăm că încrederea nu este un factor care afectează fericirea. Tabelele de frecvenţă sunt prezentate în tabelul 6.6.

Tabelul 6.6 prezintă informaţii despre cele două variabile. Respondenţii au fost rugaţi să spună dacă cred că pot avea încredere în cei mai mulţi dintre oameni sau e mai bine să ai grijă în relaţiile cu oamenii. Majoritatea aleg a doua variantă de răspuns.

Tabelul 6.6. Tabele de frecvenţă : inspectarea variabilelor înainte de analiza de contingenţă (Crosstabs)

V10 Feeling of happiness					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Very happy	205	13.6	13.7	13.7
	2 Rather happy	833	55.4	55.7	69.4
	3 Not very happy	397	26.4	26.6	96.0
	4 Not at all happy	60	4.0	4.0	100.0
	Total	1495	99.5	100.0	
Missing	-2 No answer	4	.3		
	-1 Don´t know	4	.3		
	Total	8	.5		
Total		1503	100.0		

V24 Most people can be trusted					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Most people can be trusted	115	7.7	7.7	7.7
	2 Need to be very careful	1373	91.4	92.3	100.0
	Total	1489	99.0	100.0	
Missing	-2 No answer	4	.2		
	-1 Don´t know	11	.7		
	Total	14	1.0		
Total		1503	100.0		

Nonrăspunsurile sunt definite : codurile valide sunt grupate în rândul **Valid**, iar codurile de nonrăspuns sunt grupate în rândul **Missing**. Remarcăm, aşa cum ne aşteptam, că a patra categorie de fericire, „deloc fericit”, are o frecvenţă mult mai scăzută decât celelalte. Pentru moment, obiectivul nostru de cercetare este să vedem dacă încrederea este asociată cu fericirea sau nu. Este suficient, aşadar, să am doar două categorii la variabila dependentă : fericiţi şi nefericiţi. Aşadar, folosind meniul **Transform > Recode into Different Variables**, vom crea o nouă variabilă dummy, cu numele v10rec, pornind de la V10 : codurile 1 şi 2 devin 1, fericiţi, iar codurile 3 şi 4 devin 0, nefericiţi. Mergând la încredere, observăm că majoritatea românilor nu au încredere în semenii lor, alegând varianta 2 de răspuns. Distribuţia răspunsurilor ar putea proveni din modul în care este formulat itemul : nu reuşeşte să discrimineze între indivizi. O discuţie detaliată despre acest gen de situaţii poate fi consultată în Mărginean (1982). Pe de altă parte, aceasta ar putea fi realitatea în România anului 2012. Dacă ne uităm la distribuţia răspunsurilor la această variabilă în alte ţări incluse în studiu, vom observa că arată diferit : În Australia, 48% aleg a doua variantă, în Japonia, 61%, în Noua Zeelandă, 43%, în Suedia, 38% etc. Distribuţii similare cu cea din ţara noastră sunt întâlnite în Cipru, Peru etc. Acceptăm că putem folosi itemul în analiza noastră.

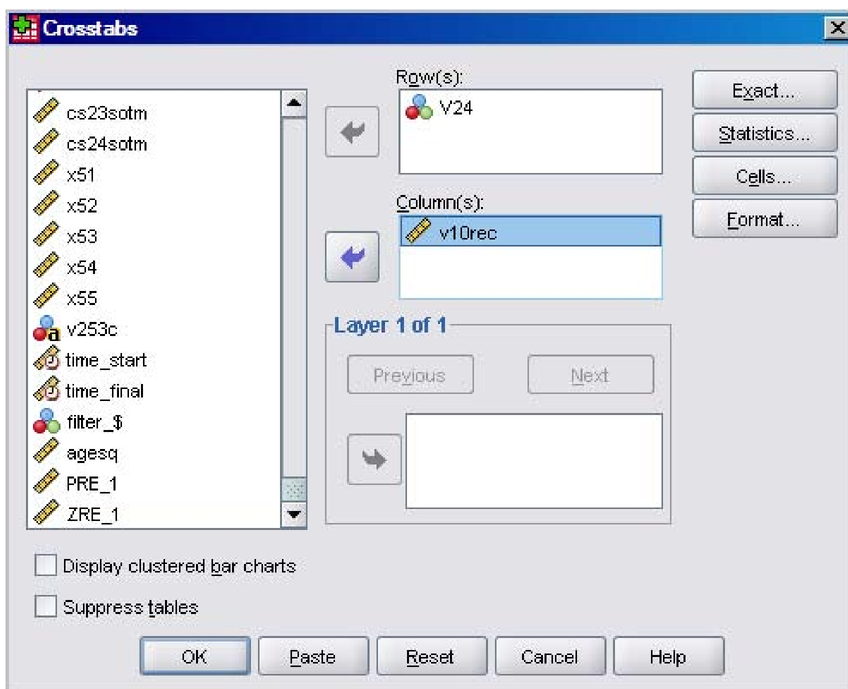
Tabelul de contingenţă este obţinut din meniul **Analyze > Descriptive Statistics > Crosstabs**. Figura 6.4a prezintă fereastra principală care se deschide prin accesarea acestui meniu.

În stânga, observăm lista de variabile din care le alegem pe cele care ne interesează şi le trecem în căsuţele din dreapta. La **Row(s)** introducem variabila care vrem să fie poziţionată pe rândurile tabelului. La **Column(s)** introducem variabila care vrem să fie poziţionată pe coloanele tabelului. Nu există o regulă cu privire la poziţionarea pe rând sau coloană. Pe rând, e preferabil să introducem variabila cu cele mai multe categorii, iar pe coloană pe cea cu cele mai puţine categorii. Astfel obţinem un tabel care va fi mai uşor de încadrat într-o coală A4 orientată portret. În exemplul nostru, această discuţie este irelevantă pentru că ambele variabile au doar două categorii de răspuns.

Dacă dorim să observăm relaţia dintre cele două variabile introduse în **Row(s)** şi **Column(s)**, în funcţie de valorile altei variabile, atunci vom utiliza **Layer 1 of 1**. De exemplu, vrem să vedem relaţia dintre încredere şi fericire, în funcţie de genul respondentului : care este relaţia pentru femei şi care este relaţia pentru bărbaţi ? Folosind butonul **Next**, care se activează după ce introducem prima variabilă în **Layer 1 of 1**, putem subdivide şi mai mult. Când folosim această opţiune, trebuie să avem destul de multe cazuri în eşantion pentru a fi relevante rezultatele.

Figura 6.4. Meniul Crosstabs

(a)



(b)

Crosstabs: Cell Display

Counts
- [x] Observed
- [ ] Expected

Percentages
- [x] Row
- [ ] Column
- [ ] Total

Residuals
- [ ] Unstandardized
- [ ] Standardized
- [x] Adjusted standardized

Noninteger Weights
- ( ) Round cell counts
- ( ) Round case weights
- ( ) Truncate cell counts
- ( ) Truncate case weights
- (•) No adjustments

Continue | Cancel | Help

(c)

Crosstabs: Statistics

- [x] Chi-square
- [ ] Correlations

Nominal
- [ ] Contingency coefficient
- [ ] Phi and Cramer's V
- [ ] Lambda
- [ ] Uncertainty coefficient

Ordinal
- [ ] Gamma
- [ ] Somers' d
- [ ] Kendall's tau-b
- [ ] Kendall's tau-c

Nominal by Interval
- [ ] Eta

- [ ] Kappa
- [ ] Risk
- [ ] McNemar

- [ ] Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Continue | Cancel | Help

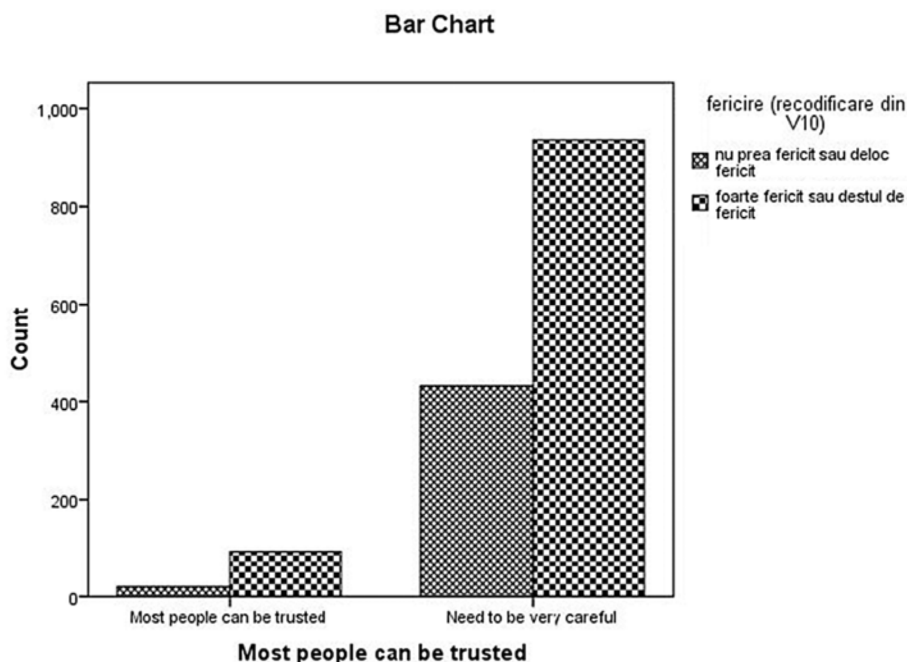
În colţul din stânga jos, observăm două opţiuni : **Display clustered bar charts** şi **Suppress tables**. Prima produce un grafic bară similar cu cel din figura 6.5a. Forma prezentată aici este modificată faţă de cea produsă prin setările implicite de către SPSS. Modificările au fost făcute dând dublu-click pe graficul rezultat în **Output**. După ce s-a deschis pentru editare, selectăm pe rând fiecare bară. Apăsăm pe meniul **Edit > Properties** din fereastra **Chart Editor**. În fereastra **Properties** (figura 6.5b) aleg tabul **Fill & Border**. În secţiunea **Color** selectez căsuţa colorată în alb şi, la **Pattern**, modelul dorit. Repet operaţiunea pentru cealaltă bară din tabel. Puteţi realiza mai multe modificări din acest meniu, în funcţie de preferinţe.

Dacă bifăm cealaltă opţiune, **Suppress tables**, atunci rularea comenzii nu va afişa tabelul de contingenţă. Dacă aţi bifat vreo opţiune în meniul care se deschide prin apăsarea butonului **Statistics**, atunci va fi afişat tabelul cu statisticile respective.

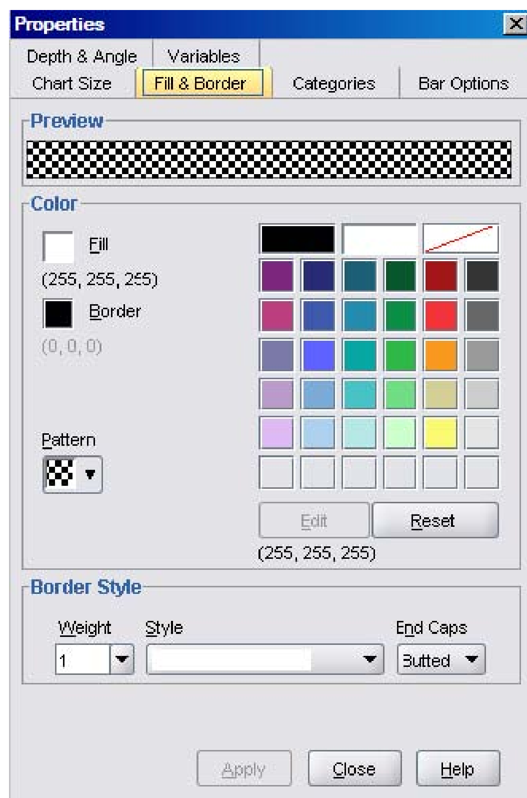
Scopul nostru principal este să vizualizăm sub formă de tabel relaţia dintre cele două variabile. Dacă după ce am introdus cele două variabile pe rând şi pe coloană (figura 6.4a) apăsăm **OK**, tabelul rezultat va conţine doar frecvenţele absolute, adică numărul de persoane care au sau nu încredere în semenii lor şi în starea acestora, de fericire sau nefericire (tabelul 6.7). 20 de persoane consideră că poţi avea încredere în cei mai mulţi oameni şi se declară nefericiţi. 94 de persoane consideră că se poate avea încredere în cei mai mulţi oameni şi se declară fericiţi.

Figura 6.5. Grafic bară obţinut folosind meniul Crosstabs

(a)



(b)



Tabelul 6.7. Tabel de contingenţă care conţine doar frecvenţe absolute (Count)

V24 Most people can be trusted * v10rec fericire (recodificare din V10) Crosstabulation				
Count				
		v10rec fericire (recodificare din V10)		Total
		0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	
V24 Most people can be trusted	1 Most people can be trusted	20	94	114
	2 Need to be very careful	432	935	1367
Total		452	1029	1481

Este destul de greu să interpretăm datele vizualizate în acest mod. De aceea trebuie să transformăm frecvenţele absolute în procente. Pentru că explicăm fericirea în funcţie de încredere, vom calcula procentele pe rând, pentru că pe rând am introdus variabila independentă. Apăsăm butonul **Cells**. Alegem, în secţiunea **Percentages**, căsuţa **Row** (figura 6.4b). Astfel, totalul de 100% va fi pe fiecare rând. Tot aici am mai făcut o modificare faţă de setările implicite: în secţiunea **Noninteger Weights**, în loc de **Round cell counts**, am bifat **No adjustments**. Baza de date pe care sunt efectuate analizele aici este ponderată, iar ponderile au valori de tipul 1.410471 sau 0.780202. Dacă nu facem modificarea, atunci când calculează statisticile, programul va rotunji sau trunchia aceste valori. Rezultatul final nu va folosi ponderile în mod corespunzător.

Tabelul de contingenţă, care include frecvenţele absolute (**Count**) şi procentele pe rând (**% within V24...**), este prezentat în tabelul 6.8. Pentru că am lăsat activă, în secţiunea **Counts**, opţiunea **Observed**, tabelul conţine atât frecvenţele absolute, cât şi procentele calculate din variabila încredere (totalurile pe rând sunt egale cu 100%). Pentru că am modificat opţiunea din secţiunea **Noninteger Weights**, frecvenţele absolute au zecimale. În raport, folosim valorile rotunjite atât la frecvenţele absolute, cât şi la procente. În terminologia procentelor alese şi calculate aici, 32% dintre cei care consideră că nu poţi avea încredere în majoritatea oamenilor se declară nefericiţi.

Tabelul este destul de dificil de citit, având prea multă informaţie. Este util să rămână vizibile doar procentele, pentru a putea detecta dacă patternul aşteptat prin ipoteza de lucru există sau nu. Putem să ne întoarcem în meniu şi să debifăm opţiunea **Observed**, lăsând doar opţiunea **Row**. Dar mai rapid ar fi să edităm tabelul în **Output**.

Tabelul 6.8. Tabel de contingenţă care conţine frecvenţe absolute şi procente pe rând

V24 Most people can be trusted * v10rec fericire (recodificare din V10) Crosstabulation					
			v10rec fericire (recodificare din V10)		Total
			0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	
V24 Most people can be trusted	1 Most people can be trusted	Count	19.926	94.167	114.093
		% within V24 Most people can be trusted	17.5%	82.5%	100.0%
	2 Need to be very careful	Count	431.848	935.166	1367.014
		% within V24 Most people can be trusted	31.6%	68.4%	100.0%
Total		Count	451.774	1029.333	1481.108
		% within V24 Most people can be trusted	30.5%	69.5%	100.0%

Aşadar, în **Output**, dăm dublu click pe tabel. Tabelul se va deschide pentru editare. În acest moment, putem modifica etichetele, putem şterge sau adăuga

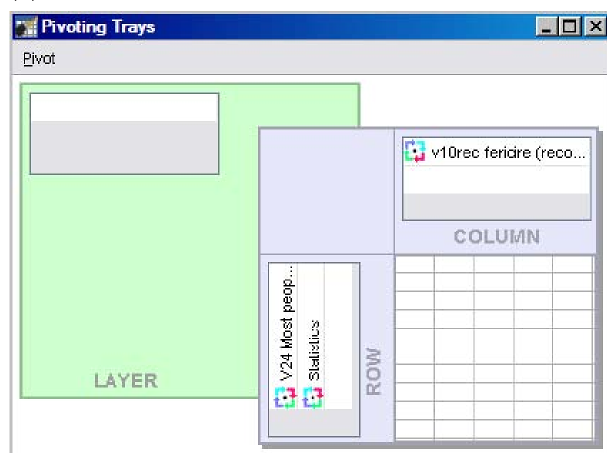
informaţie ş.a.m.d. Dar nu aceste lucruri ne interesează. Scopul nostru este să rămână vizibile doar procentele. Pentru aceasta, având tabelul deschis pentru editare, citim bara de meniuri şi observăm că au apărut câteva opţiuni noi, printre care şi **Pivot** (figura 6.6a). Meniul, înainte de dublu click, este identic cu cel din baza de date (**Data View** sau **Variable View**).

Figura 6.6. Editarea unui tabel de contingenţă în Output (Pivot)

(a)



(b)



(c)

Crosstabs

[DataSet1] C:\Document

V24 Most people can be trusted * v10rec fericire (recodificare din V10)		

V24 Most people can be

Statistics: Count

		v10rec fericire (recodificare din V10)		
		0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	Total
V24 Most people can be trusted	1 Most people can be trusted	19.926	94.167	114.093
	2 Need to be very careful	431.848	935.166	1367.014
Total		451.774	1029.333	1481.108

(d)

V24 Most people can be trusted * v10rec fericire (recodificare din V10) Crosstabulation

Statistics=% within V24 Most people can be trusted

		v10rec fericire (recodificare din V10)		Total
		0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	
V24 Most people can be trusted	1 Most people can be trusted	17.5%	82.5%	100.0%
	2 Need to be very careful	31.6%	68.4%	100.0%
Total		30.5%	69.5%	100.0%

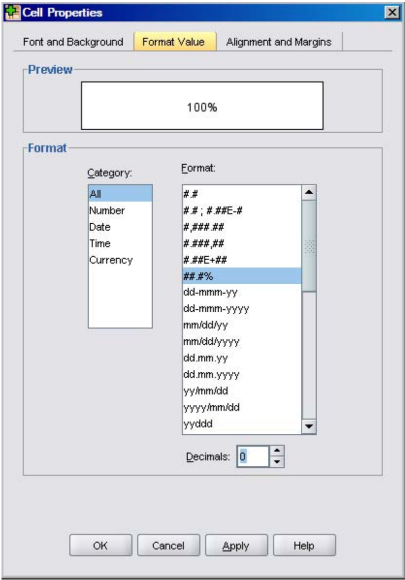
(e)

V24 Most people can be trusted * v10rec fericire (recodificare din V10) Crosstabulation

Statistics % within V24 Most people can be trusted

		v10rec fericire (recodificare din V10)		Total
		0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	
V24 Most people can be trusted	1 Most people can be trusted	17.5%	82.5%	100.0%
	2 Need to be very careful	31.6%	68.4%	100.0%
Total		30.5%	69.5%	100.0%

(f)



La **Pivot**, selectăm **Pivoting Trays** şi se deschide fereastra din figura 6.6b. Fereastra aceasta are două elemente : tabelul (fereastra din plan apropiat) şi layerul (fereastra din plan îndepărtat). În tabel, pe coloană (**COLUMN**) avem variabila v10rec, iar pe rând (**ROW**) avem variabila V24 şi statisticile calculate, mai exact frecvenţele absolute şi procentele pe rând. Ducem mouse-ul pe textul **Statistics** din **ROW** şi, ţinând apăsat, folosind procedeul drag-and-drop, tragem de acesta până în colţul din stânga sus de la **Layer**, care este alb. Când ajungem pe suprafaţa albă, săgeata mouse-ului se va transforma într-o mână. În acest moment eliberăm **Statistics**. În **Output**, observăm cum s-a modificat tabelul (figura 6.6c). În acest moment afişează doar frecvenţele absolute (**Count**). Pentru că suntem interesaţi să afişeze procentele, ducem mouse-ul pe butonul de deasupra tabelului Statistics % within V24 Most people can be trusted ▼ şi alegem **% within V24**. Modificarea este instantanee, putând citi imediat informaţiile (figura 6.6d). Pentru a închide fereastra de editare a tabelului, este suficient ca, în **Output**, să dăm click în afara lui.

Pentru că nu raportăm procentele cu virgulă, vom da iarăşi dublu click pe tabel. Selectăm toate celulele cu procente în ele (figura 6.6e) şi, în meniul care se deschide, vom selecta **Format > Cell Properties > Format value > Decimals = 0** (figura 6.6e). Apăsăm butonul **Apply** şi apoi **OK**.

Ipoteza spune că oamenii care au încredere în semenii lor sunt mai fericiţi : 83% dintre cei care au încredere în semenii lor sunt fericiţi şi 68% dintre cei care nu au încredere în semenii lor sunt fericiţi. Procentele par să susţină ideea noastră.

Înainte însă ar fi util să rulăm un test de semnificaţie. Discuţia în detaliu, despre ce sunt testele de semnificaţie, care sunt argumentele pro şi contra utilizării lor ş.a., depăşeşte scopul acestei lucrări. Cititorul este rugat să consulte lucrările dedicate acestui subiect, având în vedere importanţa pe care o au în analizele statistice. De asemenea, este rugat să înţeleagă care este relaţia cu utilizarea intervalelor de încredere pentru realizarea de inferenţe. Pentru înţelegerea corectă a acestui concept, trebuie să înţeleagă diferenţa între populaţie şi eşantion, parametru şi statistică, eşantion probabilist şi eşantion neprobabilist, trebuie înţelese concepte precum probabilitate, distribuţie de eşantionare etc. Utile în acest sens sunt lucrările scrise de Henkel (1976) şi de Mohr (1990). Din ipoteza de cercetare, sunt derivate o serie de ipoteze statistice. Un test de semnificaţie caută să verifice dacă putem să respingem ipoteza de nul. Ipoteza de nul, aşa cum sugerează numele acesteia, presupune, de exemplu, că două variabile sunt independente, adică nu au nici o relaţie, nu sunt asociate. Cercetătorul testează această ipoteză folosind un eşantion probabilist extras din populaţia de referinţă pentru studiul său. De exemplu, testăm independenţa dintre încredere şi fericire folosind un eşantion reprezentativ pentru populaţia care are 18 ani sau peste, locuieşte în România, nefiind instituţionalizată. Pornind de la acest eşantion care are, să zicem,

un volum de 1.500 de persoane, cercetătorul va face inferenţe pentru întreaga populaţie din care a fost extras. Acesta este însă unul dintre eşantioanele care puteau fi extrase folosind aceeaşi schemă de eşantionare. Dacă aplicăm aceiaşi paşi şi acelaşi algoritm, vor rezulta eşantioane care includ alte persoane decât eşantioanele extrase anterior. În ce măsură rezultatul din eşantionul nostru se datorează întâmplării ? Există o relaţie între încredere şi fericire în cadrul populaţiei ?

Pentru a răspunde la această întrebare, putem folosi testul Pearson chi-square sau, dacă dorim să îl citim în limba română, hi pătrat. Acest test presupune realizarea unui tabel de contingenţă : sunt comparate frecvenţele observate din fiecare celulă cu frecvenţele aşteptate din pură întâmplare pentru celulele respective. Calculele sunt explicate detaliat în multe lucrări de statistică, cum ar fi cea scrisă de Field (2009). Înainte de a calcula acest test, trebuie să alegem o valoare teoretică standard a nivelului de semnificaţie cu care să o comparăm pe cea calculată de program. În ştiinţele sociale, cele mai utilizate sunt 0.05, pentru un nivel de încredere de 95%, şi 0.01, pentru un nivel de încredere de 99%. Dacă alegem valoarea teoretică 0.05, iar cea calculată este mai mică decât aceasta, să zicem 0.02, atunci putem respinge ipoteza de nul a independenţei celor două variabile. Probabilitatea de a greşi spunând că încrederea şi fericirea sunt asociate este mică. Atenţie la limbaj : discuţia se poartă în termeni probabilistici. Nu putem spune : „sigur există o relaţie", ci „este mai probabil să existe decât să nu existe". În practică, când p calculat de SPSS este mai mic decât pragul teoretic utilizat, să zicem 0.05, spunem că relaţia este semnificativă statistic. Sau, şi mai scurt, că relaţia este semnificativă. Să nu confundăm însă sensul de aici cu ideea de relaţie puternică. Înseamnă doar că putem respinge ipoteza de nul, nu şi că relaţia este puternică. Înseamnă doar că probabilitatea de a greşi spunând că fericirea este asociată cu încrederea este mai mică de 0.05 sau 5%. Aceste praguri teoretice sunt relativ arbitrare. Nu există o justificare solidă teoretic pentru alegerea lor. De ce un p calculat egal cu 0.06 face relaţia nesemnificativă statistic, iar un p calculat egal cu 0.05 o face semnificativă statistic ? S-a dezvoltat o literatură alternativă pe acest subiect care merită consultată (Kline, 2004).

Testul chi-square este obţinut apăsând butonul **Statistics** : în fereastra care se deschide, bifăm **Chi-square** (figura 6.4c). Rezultatele pentru analiza asocierii dintre încredere şi fericire sunt prezentate în tabelul 6.9. Ne interesează primul rând. Coloana **Asymp. Sig. (2-sided)** conţine valoarea p calculată. Aici este egală cu 0.002. O comparăm cu valoarea teoretică 0.05, aleasă înainte de a rula analiza. În sine, testul chi-square nu ne spune mare lucru şi, dacă nu ţinem cont de anumite asumpţii ale acestuia, poate chiar să dezinformeze (Reynolds, 1984). Aici ne spune că încrederea şi fericirea sunt asociate statistic : p calculat = 0.002, valoare mai mică decât 0.05. Acest rezultat are un grad de acurateţe ridicat dacă celulele tabelului de contingenţă conţin o anumită frecvenţă aşteptată (vezi prima notă de sub tabel : **0 cells (0.0%) have expected count less than 5**). În eşantioanele cu multe unităţi, este foarte posibil ca p calculat să fie mai mic decât 0.05, chiar dacă variabilele sunt slab asociate. Argumentele teoretice pentru investigarea

acestei relaţii trebuie să fie bine gândite. O metodă empirică de verificare a acestei situaţii presupune utilizarea unei alte informaţii pe care ne-o poate calcula SPSS : calcularea valorilor reziduale ajustate (**adjusted standardized residuals**) (figura 6.4b). Acestea ne arată care celule explică asocierea dintre cele două variabile (p calculat la chi-square mai mic decât 0.05). Într-un tabel 2x2, adică dintre două variabile dihotomice, nu este prea relevant să ne uităm la aceste statistici, însă într-unul care are cel puţin o variabilă cu mai mult de două categorii se pot dovedi foarte utile în explicaţie (Field, 2009).

Tabelul 6.9. Testul Pearson chi-square : valoare şi p

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.912[a]	1	.002		
Continuity Correction[b]	9.257	1	.002		
Likelihood Ratio	10.917	1	.001		
Fisher's Exact Test				.001	.001
Linear-by-Linear Association	9.905	1	.002		
N of Valid Cases	1481				
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 34.80.					
b. Computed only for a 2x2 table					

În literatura de specialitate, starea civilă este considerată un predictor al fericirii. Unii autori susţin că implicarea într-o relaţie de cuplu creşte fericirea partenerilor (Zimmerman şi Easterlin, 2006). Folosind datele WVS 2012, putem inspecta, într-o primă fază, relaţia dintre fericire şi starea civilă. Folosesc fericirea recodificată similar cu exemplul discutat anterior. Starea civilă are trei categorii : căsătorit sau angajat într-o relaţie de cuplu ; divorţat, separat sau singur ; văduv. Cele două variabile se numesc v10rec, respectiv v57rec. Rezultatul, incluzând reziduurile ajustate, este prezentat în tabelul 6.10. Valoarea lui chi-square este 90.692 (2 grade de libertate), iar valoarea lui p calculat este mai mică decât 0.01. Relaţia dintre starea civilă şi fericire este probabilă. Mai mult, dacă ne uităm la reziduurile ajustate, aceasta este dată de fiecare tip de stare civilă. Reziduurile ajustate mai mari de 1.96, ignorând semnul, arată o relaţie semnificativă la nivel de celulă pentru un nivel de încredere de 95%. Reziduurile ajustate mai mari de 2.58, ignorând semnul, arată o relaţie semnificativă la nivel de celulă pentru un nivel de încredere de 99%. Semnele ne arată direcţia relaţiei. Reziduul −4.4 ne arată că persoanele care au o relaţie de cuplu, formalizată sau nu, nu trăiesc o stare de nefericire. În schimb, reziduul 9.5 ne arată că persoanele văduve trăiesc o stare de nefericire. Decesul partenerului reprezintă o pierdere grea în viaţa unei persoane. Trebuie investigat mai departe care sunt factorii care îi ajută pe cei divorţaţi, separaţi sau singuri să compenseze efectul pozitiv al prezenţei unui partener de viaţă. Analizele multivariate se dovedesc utile în acest sens.

Tabelul 6.10. Tabel de contingenţă cu statistici : stare civilă şi fericire

v57rec starea civila (recodificare din V57) * v10rec fericire (recodificare din V10) Crosstabulation					
		Statistics	v10rec fericire (recodificare din V10)		Total
			0 nu prea fericit sau deloc fericit	1 foarte fericit sau destul de fericit	
v57rec starea civila (recodificare din V57)	1 casatorit sau are o relatie	Count	263	720	983
		% within v57rec starea civila (recodificare din V57)	27%	73%	100%
		Adjusted Residual	-4.4	4.4	
	2 divortat, separat sau singur	Count	91	258	348
		% within v57rec starea civila (recodificare din V57)	26%	74%	100%
		Adjusted Residual	-2.1	2.1	
	3 vaduv	Count	100	58	158
		% within v57rec starea civila (recodificare din V57)	63%	37%	100%
		Adjusted Residual	9.5	-9.5	
Total		Count	454	1036	1490
		% within v57rec starea civila (recodificare din V57)	30%	70%	100%

Într-o altă analiză, am putea fi interesaţi să vedem dacă există o asociere între starea civilă şi încrederea în oameni. Aşteptarea noastră este că persoanele care au ieşit dintr-o relaţie de cuplu vor fi mai reticente în a se încrede în alte persoane. Tabelul 6.11 prezintă rezultatele analizei bivariate. Valoarea lui chi-square este 9.680 (2 grade de libertate), iar valoarea p calculată este mai mică decât 0.01. Putem respinge ipoteza de nul a independenţei celor două variabile. Inspectând reziduurile ajustate, observăm că relaţia dintre cele două variabile se datorează în principal statutului de divorţat, separat sau singur, deoarece reziduurile ajustate din dreptul acestei categorii au valori mai mari de 2.58, ignorând semnul, pe când celelalte au valori mai mici de 1.96, ignorând semnul. Aşteptarea noastră este însă confirmată parţial, deoarece aceştia consideră că se poate avea încredere în majoritatea oamenilor (reziduul ajustat = 3.0). Trebuie investigate motivele pentru care se întâmplă acest lucru. Ipoteza noastră se baza pe ideea că relaţia de cuplu încetează pentru că cel puţin unul dintre parteneri a găsit o alternativă mai bună din diferite puncte de vedere. Astfel, celălalt se va simţi trădat. Însă datele ne lărgesc orizontul. Controlând şi pentru această idee, ar trebui văzut în ce măsură persoanele care

aleg să caute un alt parteneriat sunt mai deschise la experimentare, mai permeabile la schimbare etc. Analiza multivariată se dovedeşte din nou utilă. Trebuie să ne întoarcem la teorie, să o analizăm mai atent, să vedem ce ne-a scăpat şi să construim un model explicativ pe care să îl testăm folosind o analiză care permite utilizarea simultană a mai multor variabile independente.

Tabelul 6.11. Tabel de contingenţă cu statistici : stare civilă şi încredere în oameni

v57rec starea civila (recodificare din V57) * V24 Most people can be trusted Crosstabulation					
		Statistics	V24 Most people can be trusted		Total
			1 Most people can be trusted	2 Need to be very careful	
v57rec starea civila (recodificare din V57)	1 casatorit sau are o relatie	Count	67	914	980
		% within v57rec starea civila (recodificare din V57)	7%	93%	100%
		Adjusted Residual	-1.8	1.8	
	2 divortat, separat sau singur	Count	40	309	349
		% within v57rec starea civila (recodificare din V57)	11%	89%	100%
		Adjusted Residual	3.0	-3.0	
	3 vaduv	Count	8	147	154
		% within v57rec starea civila (recodificare din V57)	5%	95%	100%
		Adjusted Residual	-1.3	1.3	
Total		Count	114	1369	1483
		% within v57rec starea civila (recodificare din V57)	8%	92%	100%

În meniul din figura 6.4c, putem alege dintre mai mulţi indicatori de asociere şi chiar de corelaţie. Aceştia sunt grupaţi în funcţie de tipul variabilelor pe care dorim să le asociem : nominale cu nominale (**Contingency coefficient**, **Phi and Cramer's V**, **Lambda**, **Uncertainty coefficient**), ordinale cu ordinale (**Gamma**, **Somers' d**, **Kendall's tau-b**, **Kendall's tau-c**), metrice cu metrice (**Correlations**) etc. Logica acestora va fi înţeleasă după parcugerea capitolului dedicat corelaţiei metrice. Spre deosebire de testul chi-square, aceştia sunt indicatori care iau o valoare într-un interval şi ne arată direcţia şi intensitatea relaţiei. O prezentare excelentă a diferenţei dintre aceştia şi a momentelor în care este potrivit să îl utilizăm pe unul sau altul a fost realizată de Chen şi Popovich (2002).

6.3. Diferenţe între medii : testul t pentru eşantioane independente şi ANOVA

Uneori suntem interesaţi să vedem dacă două grupuri au valori similare pentru o anumită caracteristică. În situaţii puţin mai complexe, am putea fi interesaţi să comparăm trei sau mai multe grupuri după o anumită caracteristică. Lucrăm cu două variabile simultan : una categorială, care dă grupurile, şi una metrică, pentru care calculăm media în cadrul fiecărui grup. Variabila categorială este cea care explică – variabilă independentă. Variabila metrică este cea explicată – variabila dependentă. Într-un studiu care analizează discriminarea femeilor pe piaţa muncii, putem fi interesaţi să comparăm salariul lunar al femeilor şi bărbaţilor care au locuri de muncă similare. Variabila de grupare va fi sexul, iar cea pentru care calculăm mediile va fi salariul lunar. Într-o cercetare de marketing, putem fi interesaţi să comparăm volumul vânzărilor, într-o anumită perioadă, pentru anvelopele de iarnă şi pentru cele de vară produse de o anumită companie. Variabila de grupare va fi tipul de anvelope, iar cea pentru care calculăm mediile va fi volumul vânzărilor.

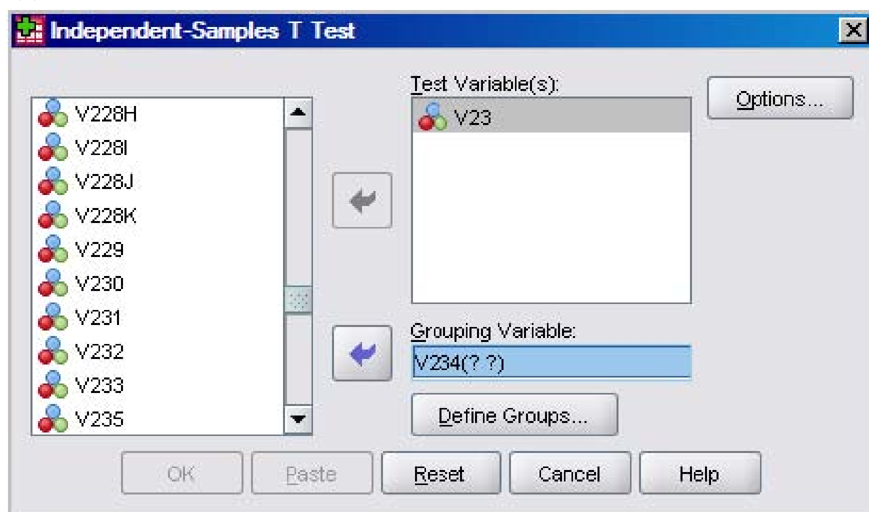
Ipoteza de nul va fi că mediile grupurilor comparate sunt egale. Dacă valoarea p calculată este mai mică decât pragul critic ales ca referinţă, 0.05 sau 0.01, atunci respingem ipoteza de nul şi considerăm plauzibilă ipoteza alternativă. Grupurile comparate diferă în ceea ce priveşte caracteristica respectivă. Salariul bărbaţilor ar putea fi mai mare decât cel al femeilor. Trebuie investigate motivele acestei situaţii. Anvelopele de vară ale companiei sunt vândute într-o cantitate mai mare decât anvelopele de iarnă. Trebuie aflat de ce se întâmplă acest lucru. Observăm că testul t pentru eşantioane independente sau analiza de varianţă (ANOVA) ne deschid căi interesante pentru explicarea unei situaţii. Însă, de regulă, cercetătorul nu se limitează la ele, ci, folosind modele explicative, aplică diferite tehnici de analiză multivariată pentru a reprezenta cât mai adecvat realitatea socială, ceva mai complexă decât aceste relaţii bivariate. ANOVA este necesară pentru că, dacă avem cel puţin trei grupuri şi aplicăm câte un test t în cazul fiecărei perechi, există posibilitatea să vedem diferenţe chiar şi acolo unde nu există. Adică respingem ipoteza de nul când nu trebuie (Henkel, 1976). Testele de semnificaţii despre care discutăm ne arată dacă diferenţele dintre mediile grupurilor există datorită variaţiilor aleatoare de la un eşantion la altul ori pentru că datele provin din populaţii în care mediile chiar sunt diferite (Iversen şi Norpoth, 1987). Pentru a fi relevantă comparaţia, grupurile trebuie să difere doar în ceea ce priveşte caracteristica presupusă a da diferenţa. Trebuie să aibă variaţii similare (Iversen şi Norpoth, 1987). Pentru verificarea acestei asumpţii, există mai multe teste de semnificaţii. SPSS ne oferă testul Levene. În funcţie de informaţia arătată de acesta, interpretăm şi rezultatul testului t pentru eşantioane independente şi ANOVA. O altă condiţie este ca variabila metrică să fie distribuită normal. Dacă

grupurile comparate au mărimi şi variaţii diferite, iar distribuţia este alungită sever, atunci este destul de probabil ca rezultul analizelor să nu fie adecvat (Agresti şi Finlay, 2008). Atunci când aceste asumpţii nu pot fi satisfăcute, ar fi util să înlocuim sau măcar să comparăm rezultatele celor două analize cu cele ale echivalentelor nonparametrice care pot fi calculate în SPSS. Dar aceasta este o altă discuţie.

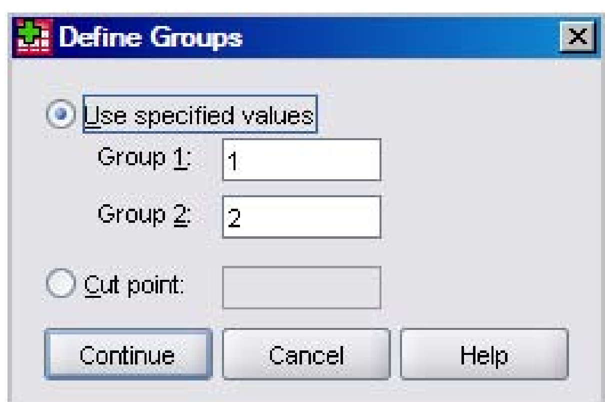
Testul t pentru eşantioane independente poate fi calculat folosind meniul **Analyze > Compare Means > Independent-Samples T Test** (figura 6.7). Fereastra care se deschide este foarte intuitivă pentru utilizator : în partea stângă avem lista de variabile din care le alegem pe cele pe care dorim să le utilizăm în analiză. În **Test Variable(s)** introducem variabila metrică pentru care dorim să calculăm media. Aici se introduce satisfacţia cu viaţa care are numele V23 în WVS 2012. În **Grouping Variable**, introducem variabila categorială care dă grupurile comparate după variabila metrică. Aici se introduce variabila V234 : „Slujba dvs. presupune să aveţi/să fi avut pe cineva în subordine ? 1. Da, 2. Nu". În figura 6.7a observăm că în dreptul numelui variabilei sunt, între paranteze, două semne de întrebare : **V234(? ?)**. SPSS solicită codurile celor două grupuri pentru care dorim să comparăm mediile satisfacţiei cu viaţa. Le aflăm dintr-un tabel de frecvenţă. Aici corespund chestionarului : 1 înseamnă că respondentul are persoane în subordine la locul de muncă, iar 2 că nu are. Pentru a le introduce, apăsăm butonul **Define Groups** (figura 6.7b). Pentru că ştim exact ce grupuri dorim să comparăm, variabila având oricum doar două coduri valide, selectăm **Use specified values** şi, la **Group 1**, respectiv **Group 2**, introducem cele două coduri (figura 6.7b). Apăsăm **Continue**, iar rezultatul este vizibil în figura 6.7c.

Figura 6.7. Meniul Independent-Samples T Test

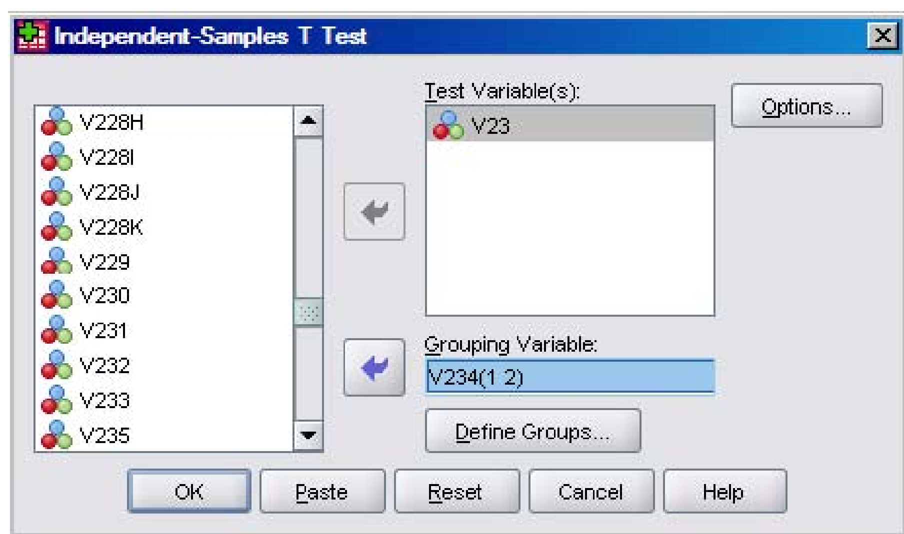
(a)



(b)



(c)



Rezultatul analizei este prezentat în tabelul 6.12. Mai întâi, sunt afişate câteva statistici descriptive. Sunt 358 de persoane care au subordonaţi, spre deosebire de cei care nu au alte persoane în subordine, care sunt în număr de 794. Cei dintâi au o medie a satisfacţiei cu viaţa egală cu 6.94, iar cei din urmă egală cu 6.63. Abaterile standard sunt apropiate ca valoare, 2.28 şi 2.35. Satisfacţia cu viaţa este măsurată pe o scală de la 1 la 10, scorurile mari indicând o satisfacţie cu viaţa mai ridicată. Următoarea figură conţine testul Levene şi testul t pentru eşantioane independente. Testul Levene ne spune că varianţele celor două grupuri sunt egale. Valoarea p calculată pentru acesta este egală cu 0.112. Fiind mai mare decât pragul teoretic de 0.05, nu putem respinge ipoteza de nul a egalităţii

varianţelor. Din acest motiv, o să citim testul t de pe rândul **Equal variances assumed**. Dacă valoarea p calculată a testului Levene ar fi fost mai mică decât 0.05, atunci am fi citit testul t de pe rândul **Equal variances not assumed**. Testul t ne spune că cele două medii sunt diferite : valoarea p calculată este egală cu 0.038, care este mai mică decât pragul teoretic de 0.05. Putem respinge ipoteza de nul a similarităţii satisfacţiei cu viaţa în rândul celor două grupuri : cu sau fără subordonaţi la locul de muncă. Din punct de vedere statistic, rezultatul ar putea fi satisfacător. Totuşi, cercetătorul nu trebuie să se mulţumească cu o abordare empiristă a realităţii sociale. Cele două medii sunt diferite prin 0.3 unităţi pe o scală de 10 puncte. Este aceasta o diferenţă de luat în seamă din punct de vedere practic ?

Tabelul 6.12. Testul t pentru eşantioane independente : output

Group Statistics					
	V234 Are you supervising someone	N	Mean	Std. Deviation	Std. Error Mean
V23 Satisfaction with your life	1 yes	358	6.94	2.281	.121
	2 no	794	6.63	2.359	.084

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
V23 Satisfaction with your life	Equal variances assumed	2.531	.112	2.077	1150	.038	.309	.149	.017	.600
	Equal variances not assumed			2.104	710.347	.036	.309	.147	.021	.597

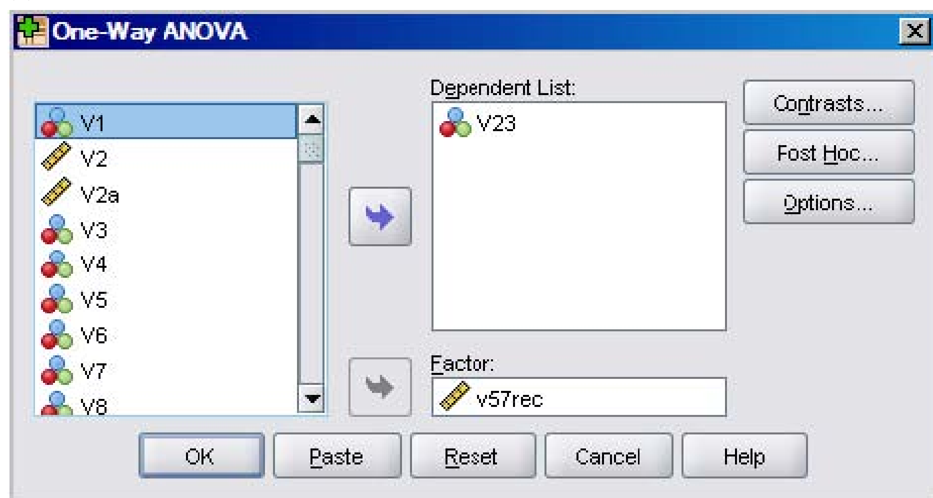
Atunci când avem mai multe grupuri ce trebuie comparate, utilizăm analiza de varianţă, prescurtată ANOVA. Nivelul satisfacţiei cu viaţa diferă în funcţie de starea civilă ? Am grupat persoanele în trei categorii : 1. Căsătorit sau trăiesc împreună cu cineva, dar nu suntem căsătoriţi ; 2. Divorţat, separat (despărţit

nelegal) sau necăsătorit şi fără a locui cu un partener ; 3. Văduv. Ipoteza de nul este că satisfacţia cu viaţa este similară pentru toate cele trei grupuri. Totuşi, noi ne aşteptăm să apară diferenţe : cei care au o relaţie ar trebui să aibă o satisfacţie cu viaţa mai mare decât în cazul celorlalţi. ANOVA este obţinută din meniul **Analyze > Compare Means > One-Way ANOVA** (figura 6.8). Fereastra este la fel de intuitivă ca la testul t pentru eşantioane independente. În partea stângă, avem lista de variabile din care le vom selecta pe cele utilizate în analiză. La **Dependent List** introducem variabila metrică, pentru care calculăm mediile. Aici se introduce satisfacţia cu viaţa care poartă numele V23. La **Factor** introducem variabila categorială, cea care dă grupurile pe care dorim să le comparăm după nivelul satisfacţiei cu viaţa. Aici se introduce starea civilă care poartă numele v57rec. Spre deosebire de meniul testului t pentru eşantioane independente, aici nu mai este nevoie să definim codurile grupurilor. Pentru a fi relevantă comparaţia, fiecare grup trebuie să aibă un număr decent de cazuri. Dacă nu se întâmplă acest lucru, atunci este utilă combinarea lor. De exemplu, primele două categorii ale stării civile sunt obţinute prin combinarea categoriilor din variabila iniţială. Pe lângă asigurarea unui număr rezonabil de cazuri în fiecare grup, am avut în vedere şi scopurile teoretice ale analizei.

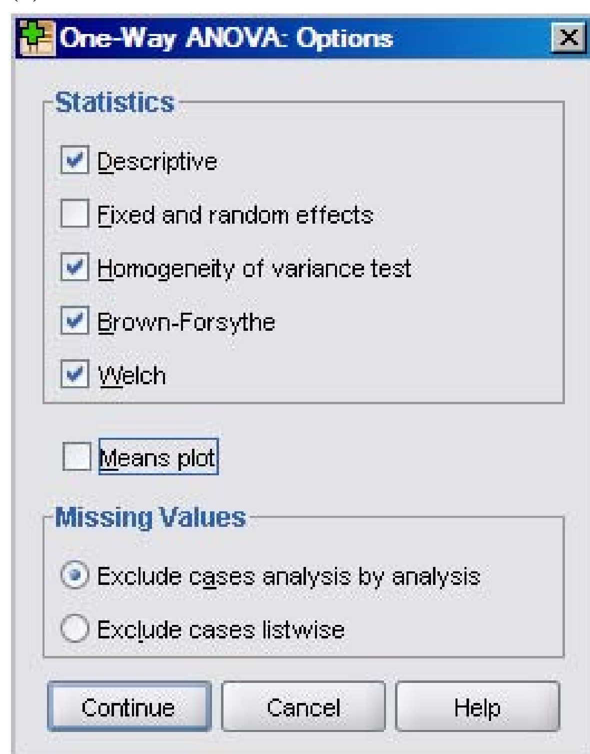
Dacă apăsăm butonul **Options**, putem alege mai multe opţiuni care vor fi afişate în **Output**. **Descriptive** ne oferă numărul de persoane din fiecare grup, media şi abaterea standard a satisfacţiei cu viaţa, dar şi intervalele de încredere în jurul mediilor şi nu numai. **Homogeneity of variance test** ne oferă testul Levene. **Brown-Forsythe** şi **Welch** sunt alternative robuste la testul F clasic specific ANOVA, atunci când asumpţia egalităţii varianţelor nu este îndeplinită. Dacă apăsăm **Continue** şi **OK**, obţinem rezultatul din tabelul 6.13a. Testul Levene ne spune că varianţele nu sunt egale : p calculat este mai mic decât 0.01 (coloana **Sig**). Acest lucru îl intuiam după ce am comparat abaterile standard. Putem să mai consultăm forma distribuţiei satisfacţiei cu viaţa pentru cele trei grupuri realizând câte un grafic bară pentru fiecare dintre ele. Dacă pentru cei care au o relaţie şi cei care sunt divorţaţi, separaţi sau singuri distribuţiile au aproximativ aceeaşi formă, aceasta arată destul de diferit pentru văduvi. Testul **F**, din tabelul **ANOVA**, ne spune că cel puţin două dintre grupurile comparate au o satisfacţie cu viaţa diferită. Pentru că varianţele sunt inegale, am preferat să consult şi alternativele **Welch** şi **Brown-Forsythe** care, de data aceasta, oferă acelaşi rezultat ca şi testul **F**. Pentru a afla care grupuri diferă şi în ce fel, trebuie să folosim metoda comparaţiilor multiple sau testele post-hoc. Acestea pot fi accesate apăsând butonul **Post Hoc**. Aceste teste, fiecare cu avantajele şi dezavantajele sale, sunt alese în funcţie de rezultatul testului Levene. Aici, pentru că varianţele nu sunt egale, alegem unul dintre testele din secţiunea **Equal Variances Not Assumed** (figura 6.8c). Rezultatele sunt prezentate în tabelul 6.13b.

Figura 6.8. Meniul One-Way ANOVA. Analiza de varianţă

(a)



(b)



(c)

One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

LSD | S-N-K | Waller-Duncan
Type I/Type II Error Ratio: 100
Bonferroni | Tukey
Sidak | Tukey's-b | Dunnett
Control Category : Last
Scheffe | Duncan
R-E-G-W F | Hochberg's GT2
R-E-G-W Q | Gabriel
Test: 2-sided | < Control | > Control

Equal Variances Not Assumed

Tamhane's T2 | Dunnett's T3 | Games-Howell | Dunnett's C

Significance level: 0.05

Continue | Cancel | Help

Tabelul 6.13. Rezultate ale analizei de varianţă

(a)

Test of Homogeneity of Variances			
V23 Satisfaction with your life			
Levene Statistic	df1	df2	Sig.
13.805	2	1482	.000

ANOVA					
V23 Satisfaction with your life					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	284.112	2	142.056	25.670	.000
Within Groups	8201.352	1482	5.534		
Total	8485.464	1484			

Robust Tests of Equality of Means				
V23 Satisfaction with your life				
	Statistic[a]	df1	df2	Sig.
Welch	18.075	2	365.203	.000
Brown-Forsythe	21.310	2	451.001	.000

a. Asymptotically F distributed.

(b)

Multiple Comparisons						
V23 Satisfaction with your life Tamhane						
(I) v57rec starea civila (recodificare din V57)	(J) v57rec starea civila (recodificare din V57)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 casatorit sau are o relatie	2 divortat, separat sau singur	.015	.146	.999	-.34	.37
	3 vaduv	1.420*	.238	.000	.85	1.99
2 divortat, separat sau singur	1 casatorit sau are o relatie	-.015	.146	.999	-.37	.34
	3 vaduv	1.405*	.260	.000	.78	2.03
3 vaduv	1 casatorit sau are o relatie	-1.420*	.238	.000	-1.99	-.85
	2 divortat, separat sau singur	-1.405*	.260	.000	-2.03	-.78
*. The mean difference is significant at the 0.05 level.						

Mai întâi, consultăm coloana **Sig** care conţine valorile p calculate. Observăm diferenţe semnificative statistic (p calculat < 0.05) între grupurile „căsătorit sau are o relaţie” şi „văduv”, respectiv „divorţat, separat sau singur” şi „văduv”. Ipoteza de lucru se confirmă parţial, pentru că nu observăm o diferenţă între „căsătorit sau are o relaţie” şi „divorţat, separat sau singur”. Apoi consultăm coloana **Mean Difference (I-J)**. Aceasta ne spune cu cât diferă mediile grupurilor comparate. Literele **I** şi **J** desemnează prima, respectiv a doua coloană din tabel. De exemplu, diferenţa dintre media satisfacţiei cu viaţa a celor căsătoriţi sau care au o relaţie şi media văduvilor este de 1.42 unităţi. Cei dintâi au media 6.85, iar cei din urmă au media 5.43. Pasul următor este căutarea acelor factori care fac, de exemplu, ca persoanele divorţate, separate sau singure să fie mai satisfăcute cu viaţa decât cele văduve.

6.4. Două grafice uzuale în descrierea datelor

Graficele pot fi folosite în două scopuri. Un scop, pur operaţional, este vizualizarea datelor altfel decât sub formă de tabele în timpul activităţii de analiză. Celălalt scop este prezentarea informaţiei în lucrările noastre într-un mod mai intuitiv decât sub formă de tabele sau descriere în cuvinte. A face un grafic bun nu este atât de intuitiv tot timpul.

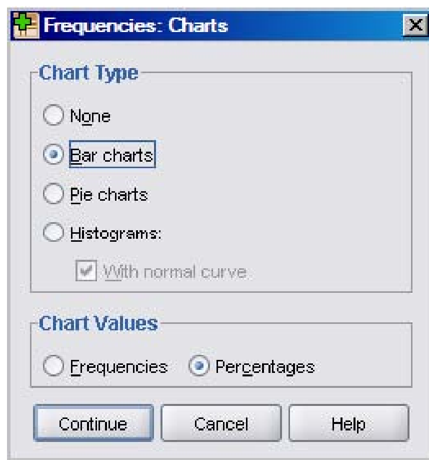
Good şi Hardin (2012) oferă câteva reguli pentru cei ce doresc să utilizeze grafice pentru prezentarea datelor. Prezentăm o listă adaptată după aceşti autori :

- Realizaţi grafice 2D. A treia dimensiune trebuie folosită doar dacă există.
- Folosiţi bare şi evitaţi formele geometrice speciale cum ar fi conul, cilindrul etc. De asemenea, evitaţi umbrele generate de bare.
- Includeţi valorile pe bare. Dacă sunt prea multe bare, includeţi valorile măcar la extreme şi la o categorie de interes major. În cazul în care comparăm salariul minim din diferite ţări europene, evidenţiem ţara cu salariul minim şi pe cea cu salariul maxim, dar şi valoarea specifică României.
- Includeţi etichete care clarifică elementele din grafic. Dacă sunt prea multe etichete, atunci alegeţi cu atenţie unele care evidenţiază ideea centrală a graficului. Nu suprapuneţi etichetele cu elementele esenţiale ale graficului.
- Evitaţi spaţiile goale de dimensiuni mari în grafice. Ajustaţi scala variabilelor astfel încât să reflecte amplitudinea din date nu pe cea ideală.
- Utilizaţi graficele în acord cu proprietăţile variabilelor.

Această listă este doar un început. Poate fi îmbogăţită şi adaptată în funcţie de informaţia care se vrea transmisă şi graficul ales în acest sens.

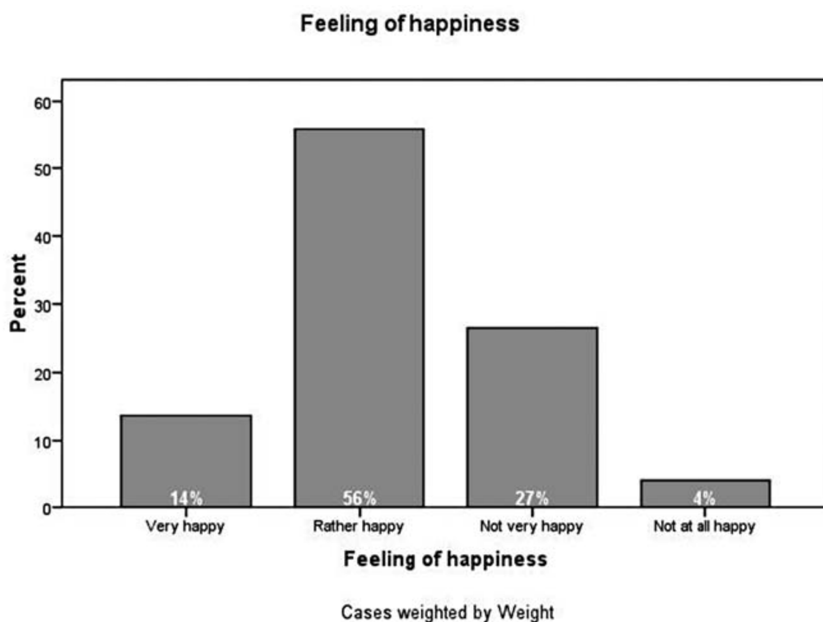
SPSS are o caracteristică utilă mai ales pentru utilizatorii novici : ne permite să realizăm grafice atât din meniurile unor analize, cât şi din meniul dedicat special acestui lucru. Pentru începători, recomand utilizarea primei variante. De exemplu, meniul **Frequencies** ne oferă posibilitatea realizării a trei grafice : radial (**pie**), bară (**bar chart**) şi histogramă (**histogram**). Intraţi în meniu şi apăsaţi butonul **Charts**. În fereastra care se deschide (figura 6.9) trebuie doar să selectăm tipul de grafic care ne interesează. Aici am ales să facem un grafic bară, axa Oy evidenţiind procentele. Folosim procentele pentru că acestea au mai mult sens atunci când le citim decât frecvenţele absolute.

Figura 6.9. Meniul Frequencies, Charts



Graficul radial nu este un instrument bun de vizualizare a datelor. Diferenţele dintre secţiunile graficului pot fi atât de mici, încât devine necesară suprapunerea valorilor peste ele. Deja oferim informaţie redundantă. O soluţie mult mai bună este graficul bară. Un exemplu este prezentat în figura 6.10. Pe axa Ox sunt cele patru niveluri de fericire. Axa Oy ne spune procentul celor care aleg o categorie sau alta.

Figura 6.10. Grafic bară



Valorile de pe fiecară bară le-am adăugat ulterior. Am dat dublu click pe grafic, acţiune în urma căreia se deschide **Chart Editor**. Mergem în meniul **Elements > Show Data Labels**. În fereastra care se deschide, selectăm tabul **Data Value Labels**, **Label Position** şi apoi **Custom**, bifând poziţia dorită (figura 6.11a).

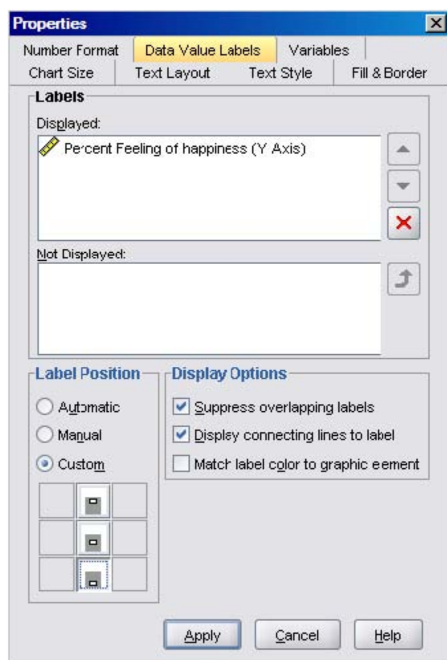
Tot aici putem modifica numărul de zecimale, tipul textului şi altele. Selectăm tabul **Number Format** şi facem următoarele modificări : **Decimal Places = 0** şi **Trailling Characters = %** (figura 6.11b). Apăsăm **Apply**, **Close** şi închidem fereastra de editare a graficului. Evident, putem face şi alte modificări în aceste ferestre.

Aici, vedem rapid că majoritatea românilor erau destul de fericiţi în 2012 conform WVS.

Când vrem să reprezentăm grafic o variabilă metrică cu multe categorii, în locul graficului bară alegem histograma (figura 6.12). Mergem în **Frequencies**, apăsăm **Charts**, şi selectăm **Histogram**. Comparând cele două histograme, observăm o variaţie mai mare în rândul femeilor în ceea ce priveşte timpul petrecut cu îngrijirea copiilor, bătrânilor sau celor bolnavi. Graficele sugerează, de asemenea, un timp mai mare petrecut de femei cu acest gen de activităţi.

Figura 6.11. Editarea valorilor de pe bare, Properties

(a)



(b)

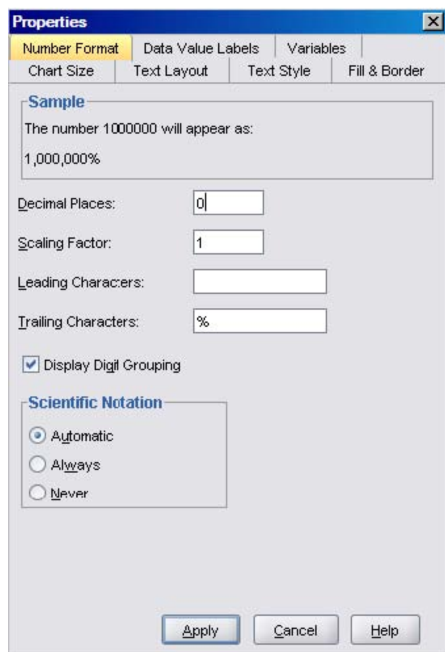
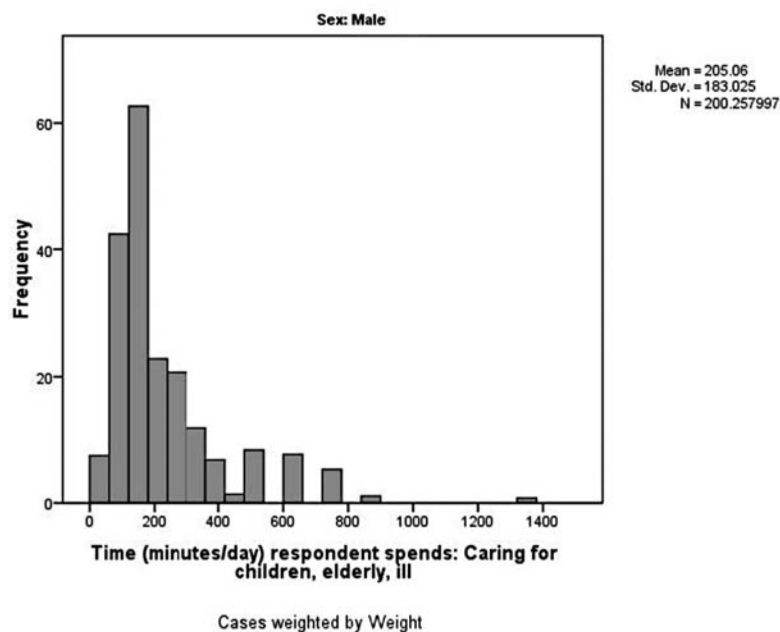
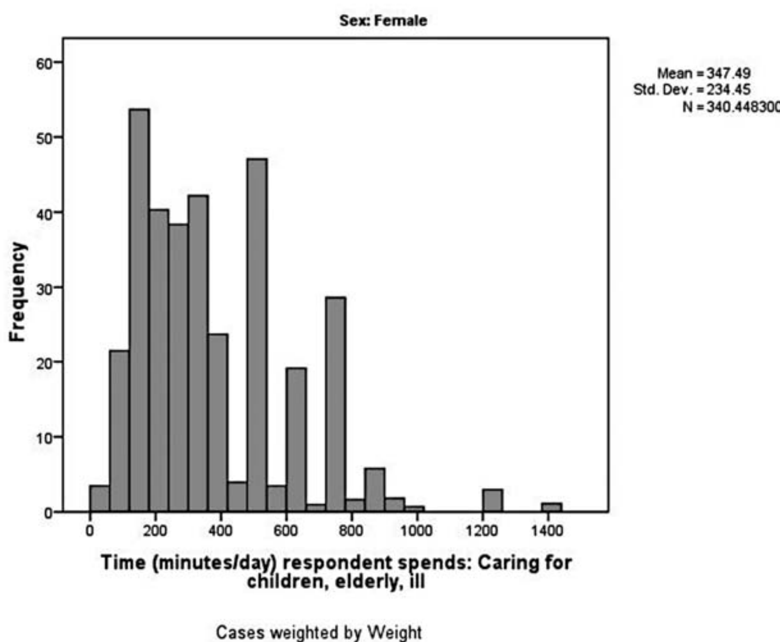


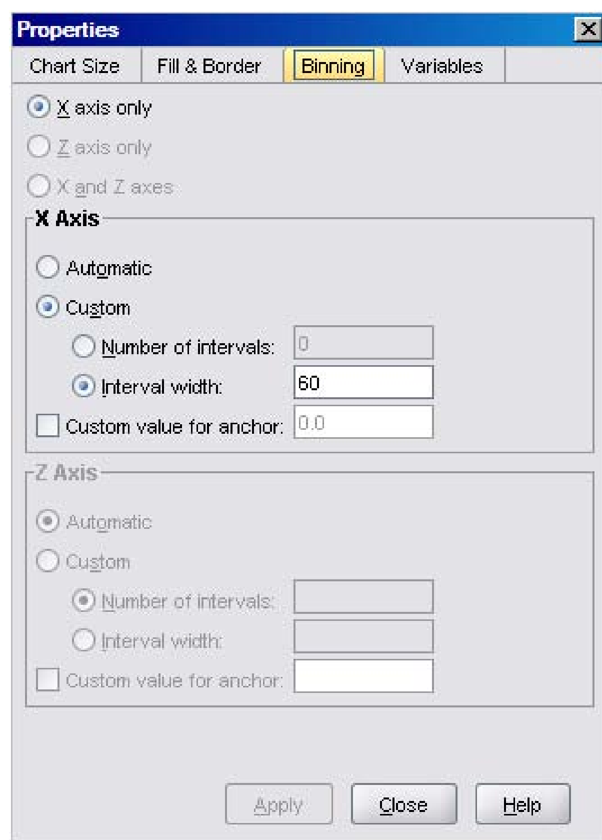
Figura 6.12. Histograme





Histogramele trebuie utilizate cu precauţie. De exemplu, aici, am modificat mărimea intervalului folosit pentru reprezentarea grafică astfel încât să fie egal cu 60 de minute. Acest lucru se poate face dând, în **Output**, dublu click pe grafic. Se deschide **Chart Editor**. În interiorul acestuia, dăm dublu click pe barele histogramei şi se deschide fereastra **Properties**. Selectăm tabul **Binning** (figura 6.13) şi, în secţiunea **X Axis**, selectăm **Custom**, iar la **Interval width** introducem valoarea dorită.

Figura 6.13. Chart Editor, Properties pentru histogramă



Există mai multe lucrări care prezintă principiile reprezentării grafice corecte. Dintre acestea le pot aminti pe cele elaborate de Chambers şi colaboratorii (1983), Jacoby (1997, 1998) şi Tufte (2001).

6.5. Exerciţii

Pentru aceste exerciţii, utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultat(ă) în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a *Grupului Românesc pentru Studiul Valorilor Sociale* (http : //www.romanianvalues.ro).

1. Deschideţi chestionarul WVS 2012. Calculaţi indicatorii tendinţei centrale şi variaţiei corespunzători pentru variabilele de pe paginile cu număr impar.
2. Exportaţi tabelele în Excel şi editaţi-le pentru includerea într-un raport.
3. Pentru variabilele de la exerciţiul anterior, realizaţi câte un grafic care să reflecte cât mai bine informaţia. Editaţi aceste grafice astfel încât să poată fi folosite într-un material tipărit monocrom (alb-negru).
4. Găsiţi în baza de date variabila care măsoară fericirea (nu satisfacţia cu viaţa). Elaboraţi o listă cu zece variabile nominale, diferite faţă de cele folosite în textul acestui capitol, care credeţi că influenţează fericirea. Realizaţi o listă de ipoteze în care fericirea este variabila explicată.
5. Realizaţi zece tabele de contingenţă în care testaţi ipotezele notate la exerciţiul anterior. Scrieţi un scurt raport de o pagină în care descrieţi ce aţi aflat, folosind valorile reziduale ajustate.
6. Realizaţi un profil al părinţilor care sunt predispuşi să transmită copiilor lor valoarea „imaginaţie". Profilul trebuie să conţină cinci variabile explicative nominale sau ordinale din chestionarul WVS 2012.
7. Testaţi profilul folosind tabele de contingenţă cu valori reziduale ajustate.
8. Creaţi o variabilă nouă care reflectă statutul de membru activ în organizaţii voluntare. Verificaţi dacă media satisfacţiei cu viaţa este diferită pentru membrii activi faţă de cei care sunt membri inactivi sau nu sunt membri.
9. Creaţi o variabilă nouă care reflectă statutul de membru activ, membru inactiv şi nonmembru în organizaţii voluntare. Verificaţi dacă media satisfacţiei cu viaţa este diferită pentru aceste trei categorii.
10. Creaţi o variabilă nouă care să reflecte intoleranţa faţă de grupuri marginale sau minorităţi (persoane dependente de droguri, persoane de rasă diferită de a dumneavoastră etc.). Aceasta trebuie să folosească setul de variabile v36-v44 din chestionarul WVS 2012. Variabila nou-creată trebuie să reprezinte numărul de grupuri pe care o persoană nu i-ar dori ca vecini. Verificaţi dacă numărul este mai mic pentru cei care fac voluntariat activ decât pentru cei care nu fac voluntariat. Verificaţi dacă există diferenţe în ceea ce priveşte numărul de categorii alese între cei care fac voluntariat activ, cei care nu fac voluntariat activ, respectiv cei care nu fac voluntariat deloc.

7. Explorarea datelor : asumpţii

Dacă două persoane au aceleaşi caracteristici, dar diferă în funcţie de venitul lunar câştigat, care va fi mai satisfăcută cu viaţa : cea cu un venit mai mare ? În oraşele mari sunt mai multe persoane fericite decât în oraşele mici ? Persoanele care au emigrat la o vârstă mai înaintată se adaptează mai uşor la modul de viaţă din ţara de destinaţie ? Studenţii care participă la activităţi de voluntariat au o şansă mai mare în a găsi un loc de muncă apropiat de aşteptările şi dorinţele lor ?

Un cercetător organizat, înainte de a trece la elaborarea chestionarului, trebuie să anticipeze ce fel de analize solicită întrebările sale de cercetare. Deşi pare contraintuitiv, punând căruţa înaintea calului, în practică nu este chiar aşa. Degeaba formulezi întrebările şi alegi variantele de răspuns dacă nu eşti conştient de calităţile psihometrice pe care acestea le au. Când primeşte datele din teren şi ai în faţă baza de date, începi să te întrebi ce poţi face cu variabilele avute la dispoziţie şi, dacă nu te-ai gândit dinainte la corespondenţa cu tehnicile statistice, răspunsul s-ar putea să nu îţi placă. Să presupunem că, măcar parţial, cercetătorul a formulat întrebările din chestionar în acord cu cerinţele statistice ale obiectivelor de cercetare. Înainte de a trece la analiza propriu-zisă a datelor, adică la aplicarea tehnicii statistice prin care acestea răspund la întrebarea de cercetare, va trebui să inspecteze variabilele univariat, bivariat sau chiar multivariat. Când explorăm datele, căutăm să înţelegem dacă variabilele au suficientă variaţie, dacă există cazuri extreme, cum arată distribuţia acestora etc.

Pentru înţelegerea unor analize descrise aici, cititorul ar trebui să aibă cunoştinţe minime despre testele de semnificaţie. Câteva noţiuni elementare din această zonă a statisticii au fost prezentate în capitolul 6. Totuşi, ar trebui să suplimentaţi această lucrare cu un manual de statistică. Am sugerat câteva astfel de lucrări în volumul de faţă.

În acest capitol, vom vorbi despre statistici şi grafice care ne ajută să decidem dacă există cazuri extreme, care este forma distribuţiei, dacă între două variabile există o relaţie liniară etc. Acestea ne ajută să ne cunoaştem datele înainte de a le folosi la calcularea unor statistici cum ar fi media sau abaterea standard. Analiştii trebuie să fie sceptici în legătură cu calitatea datelor lor. Acest scepticism previne erorile în interpretarea substanţială.

7.1. Distribuţia unei variabile

Măsurarea este procesul prin care cercetătorul operaţionalizează conceptele cu care lucrează, construieşte instrumentele prin care culege date despre fenomenele pe care le cuprinde şi verifică validitatea şi fidelitatea acestora. Prin măsurare, de exemplu, ajungem la scalele compuse pe care le folosim frecvent în chestionarele noastre. Unul dintre criteriile după care cercetătorul decide care sunt itemii pe care îi va păstra în scala finală este cel al irelevanţei (Mărginean, 1982). Pe scurt, acest criteriu presupune eliminarea, din analizele prin care căutăm răspunsurile la întrebarea de cercetare, a itemilor cu care toţi respondenţii sau o majoritate covârşitoare dintre aceştia sunt de acord sau nu. Pentru variabilele categoriale este relativ simplu să aflăm această informaţie. Putem realiza un tabel de frecvenţă şi inspecta distribuţia procentelor (tabelul 7.1).

Tabelul 7.1. Tabel de frecvenţă : verificarea variaţiei variabilelor categoriale

V219 Information source: TV news					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Daily	1235	82.1	82.3	82.3
	2 Weekly	137	9.1	9.1	91.4
	3 Monthly	19	1.3	1.3	92.6
	4 Less than monthly	81	5.4	5.4	98.0
	5 Never	30	2.0	2.0	100.0
	Total	1501	99.9	100.0	
Missing	-2 No answer	1	.0		
	-1 Don´t know	1	.1		
	Total	2	.1		
Total		1503	100.0		

Conform datelor WVS 2012, 82% dintre români foloseau, zilnic, ştirile prezentate la televizor ca sursă de informare despre ce se petrece în ţară şi în lume. Aceasta este o informaţie utilă şi interesantă despre comportamentul de informare al românilor şi nu numai. Are însă suficientă variaţie această variabilă dacă dorim să o includem într-o analiză multivariată ? Răspunsul nu este simplu de oferit. Dacă cercetătorul are argumente teoretice solide, poate decide să o folosească ca atare sau poate considera că ar fi mai util să o recodifice : televizorul este, probabil, principala sursă de informaţie şi divertisment, fiind accesibil atât în ceea ce priveşte costurile, cât şi referitor la dificultatea conţinuturilor prezentate. Presiunea timpului solicită formate scurte, concentrate, cu mesaje transmise în forme inteligibile pentru mase mari de privitori. Din acest

motiv, orice frecvenţă de utilizare, în afara celei zilnice, ar putea fi considerată aparte. Cei care se informează mai rar decât zilnic de la ştirile televizate au, probabil, caracteristici diferite faţă de ceilalţi. Viaţa socială este atât de complexă, încât decizia de a utiliza într-o formă sau alta această variabilă depinde de mulţi factori.

Să nu înţelegeţi că procesul de explorare a datelor are ca unic scop găsirea problemelor. În primul rând, dorim să ne familiarizăm cu datele. Apoi, dorim să vedem dacă sunt modificări pe care trebuie să le aducem variabilelor pentru a utiliza cât mai multă informaţie culeasă prin chestionar.

Agresti şi Finlay (2008) consideră distribuţia normală (cea care are formă de clopot) ca fiind cea mai importantă pentru analiza statistică deoarece aproximează destul de bine forma multor variabile din viaţa reală. Pentru a înţelege proprietăţile acestei distribuţii şi de ce este importantă, trebuie să consultaţi capitolele dedicate acestui subiect din lucrarea citată sau din Agresti şi Franklin (2013). Deşi, în esenţă, cele două lucrări prezintă aceeaşi informaţie, cea din urmă are o formă de prezentare grafică mai prietenoasă. În acest moment, este suficient să reţinem că multe dintre analizele inferenţiale aplicate în mod obişnuit în ştiinţele sociale folosesc această distribuţie. Concepte importante asociate sunt distribuţia de eşantionare, eroarea standard sau teorema limită centrală. Toate sunt tratate comprehensiv în lucrările citate.

Testele parametrice asumă distribuţia normală. De aceea, înainte de a rula şi interpreta astfel de teste, trebuie să verificăm dacă variabilele au sau nu o distribuţie aproximativ normală. Dacă nu au, atunci putem utiliza testele nonparametrice echivalente (Cramer şi Howitt, 2004). Printre analizele pentru care verificarea asumpţiei normalităţii distribuţiei este importantă se numără analiza de corelaţie Pearson, ANOVA, testele t şi regresia multiplă (de Vaus, 2002). Această discuţie se aplică variabilelor metrice. În practică, convenţional, se acceptă şi pentru variabilele ordinale (scală tip Likert).

Putem verifica dacă această asumpţie este îndeplinită folosind grafice sau calculând diferiţi indicatori şi teste statistice. Mai întâi discutăm despre metodele grafice.

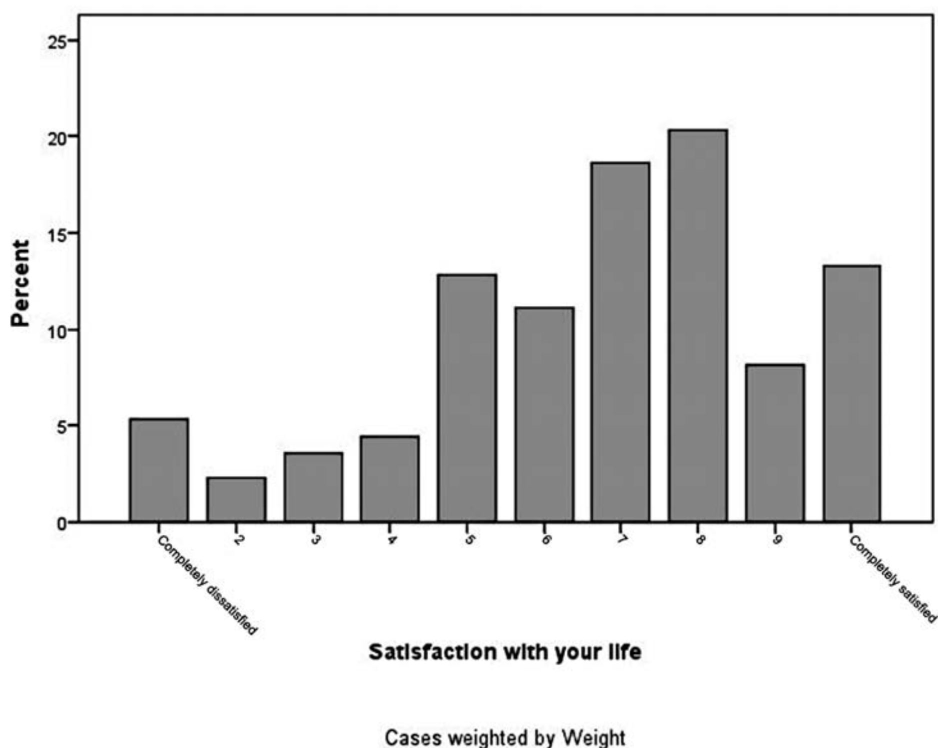
În figura 7.1 este prezentată distribuţia satisfacţiei cu viaţa pentru români în 2012 conform World Values Survey.

Acest grafic a fost obţinut din meniul **Analyze > Descriptive statistics > Freqencies > Charts**. A fost editat aplicând paşii deja discutaţi în alte locuri din acest volum.

Distribuţia satisfacţiei cu viaţa, o variabilă măsurată pe o scală de la 1 la 10, unde 1 înseamnă „total nemulţumit" şi 10 „total mulţumit", este alungită la stânga. Majoritatea românilor se poziţionează în partea pozitivă a scalei. Distribuţia se abate de la normalitate. Acest lucru nu este neapărat rău, pentru că satisfacţia cu viaţa, în realitate, are o distribuţie de acest gen (Cummins, 2003). Ceea ce

îngrijorează, gândindu-ne la validitatea rezultatelor analizelor statistice, sunt frecvenţele neaşteptat de mari pentru categoriile 1 şi 10, cei total nemulţumiţi sau mulţumiţi cu viaţa lor în general. Înainte de a calcula medii sau coeficienţi de corelaţie, trebuie să înţelegem de ce apar aceste două abateri de la normalitate. O altă problemă importantă, care este mai puţin vizibilă aici, apare atunci când există mai multe vârfuri şi goluri între aceste vârfuri. Distribuţia, în această situaţie, apare ca şi când ar fi formată din mai multe distribuţii mici. Calcularea mediei sau medianei nu ar avea foarte mult sens în această situaţie. Hartwig şi Dearing (1979) recomandă să creăm variabile separate din cea iniţială şi să le utilizăm ca atare.

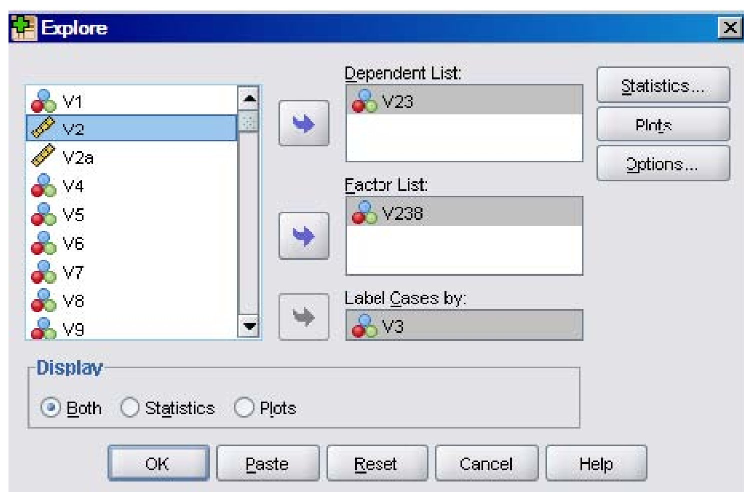
Figura 7.1. Grafic bară pentru verificarea asumpţiei de normalitate a distribuţiei



Un grafic care oferă informaţii similare cu histograma, dar care adaugă şi altele noi, este box-plot-ul. Acesta poate fi obţinut din **Analyze > Descriptive statistics > Explore > Plots > Box-plot = Factor levels together** (figura 7.2).

Figura 7.2. Meniul Explore

(a)

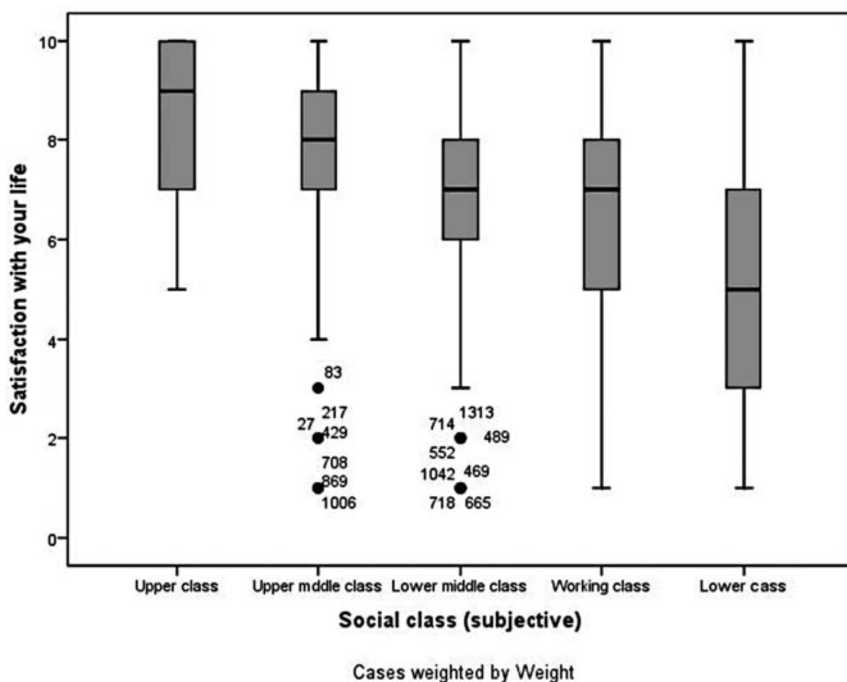


(b)



În WVS 2012, respondenţii au fost rugaţi să se poziţioneze în ierarhia socială într-una dintre poziţiile : „clasa de jos", „clasa muncitoare", „în partea de jos a clasei mijlocii", „în partea de sus a clasei mijlocii", „în clasa de sus". În figura 7.3 este prezentată distribuţia satisfacţiei cu viaţa pentru fiecare dintre aceste poziţii sociale.

Figura 7.3. Box-plot : distribuţia satisfacţiei cu viaţa în funcţie de poziţia socială subiectivă



Revenind puţin la meniul din care a fost obţinut box-plot-ul, observăm următoarele (figura 7.2a) :

- Presupunem că o variabilă, aici satisfacţia cu viaţa (V23), variază în funcţie de o alta, aici autopoziţionarea pe o scală a poziţiei sociale (V238). De aceea, variabila pentru care sunt calculate statisticile (medie, mediană, abatere standard etc.) va fi introdusă în câmpul **Dependent List**, iar variabila care dă grupurile pentru care sunt realizate comparaţiile va fi introdusă în câmpul **Factor List**. Denumirea **factor** atribuită unei variabile categoriale care distinge între anumite grupuri a mai fost întâlnită la analiza de varianţă.
- Pentru a putea identifica mai uşor eventualele cazuri cu „probleme", le etichetăm folosind variabila de identificare care nu trebuie să lipsească din nici o bază de date. Aici, această variabilă, care conţine ID-uri unice pentru fiecare respondent, este **V3**.
- Pentru a nu încărca outputul cu multe informaţii, în prima fază, putem bifa doar opţiunea **Plots** în secţiunea **Display**. Dacă lăsăm bifată opţiunea **Both**, atunci în output vor fi afişate şi statisticile produse de meniu. Aici acestea vor fi calculate pentru fiecare poziţie socială. Am putea lua în considerare suprimarea temporară a acestora.

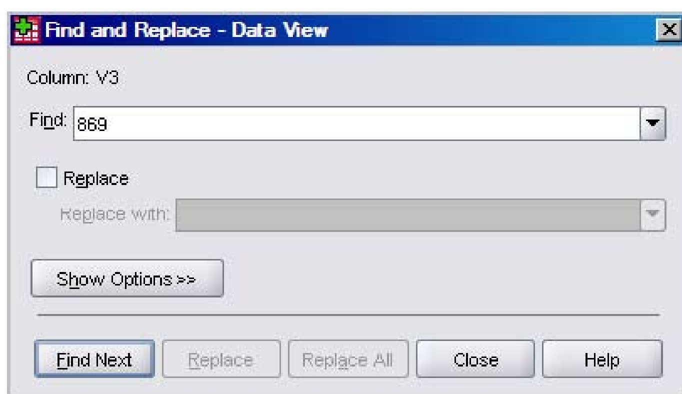
Spuneam că box-plot-ul ne oferă câteva informaţii mai puţin evidente din histogramă. Comparând poziţiile liniei îngroşate din interiorul cutiilor (mediana), observăm că aceasta ia valori din ce în ce mai mici pe măsură ce persoanele se autopoziţionează în clase sociale aflate pe trepte din ce în ce mai joase ale ierarhiei sociale. În cazul de faţă trebuie să fim precauţi cu interpretarea valorilor pentru „clasa de sus”, deoarece în această categorie sunt doar 17 persoane în eşantion. Apoi observăm că, mergând de la clasa de sus în jos, creşte şi variaţia satisfacţiei cu viaţa (cutia este mai lungă, deci abaterea intercuartilă este mai mare). Precauţia cu privire la numărul de cazuri din clasa de sus capătă şi mai mult sens aici, pentru că ne aşteptam să vedem variaţia cea mai restrânsă în această categorie. Continuăm cu citirea graficului. În clasa muncitoare, distribuţia este alungită la stânga (mediana se află înspre capătul de sus, cuartila 3, a cutiei). În cele două secţiuni ale clasei mijlocii, există cazuri extreme (outlieri) simbolizate prin cerculeţe. SPSS identifică două tipuri de cazuri extreme : cele discutate şi cele care se află foarte departe în distribuţie, reprezentate cu steluţe. Atunci când încercăm să remediem problema cazurilor extreme, întotdeauna începem cu steluţele.

Tabachnick şi Fidell (2007) oferă mai multe soluţii pentru gestionarea cazurilor extreme. Prima soluţie, cea radicală, este scoaterea din analize a persoanei sau a persoanelor respective. A doua soluţie, cea care caută să maximizeze utilizarea datelor aflate la dispoziţie, presupune aplicarea unor transformări variabilei care conţine cazurile extreme. O astfel de transformare poate fi obţinută, de exemplu, prin logaritmarea variabilei cu cazuri extreme. Cazurile vor rămâne în baza de date, dar influenţa lor va fi diminuată considerabil. O altă metodă constă în intervenţii directe asupra cazurilor extreme : valoarea extremă este recodificată în jos sau în sus. De exemplu, dacă salariul din ultima lună are valoarea extremă de 15.000 de lei, iar următoarea valoare, care nu este caz extrem, este 5.500 de lei, atunci putem recodifica în 5.600 de lei sau altă valoare aleasă în funcţie de distribuţia celorlalte valori. Soluţiile nu sunt simplu de ales. Lucrurile se complică şi mai mult dacă reţinem că discuţia, până în acest punct, a fost despre cazurile extreme univariate ignorându-le pe cele multivariate (o persoană care a absolvit facultatea, are 24 de ani şi la primul loc de muncă primeşte un salariu lunar de 10.000 de lei). Aceiaşi autori atrag atenţia că soluţiile enunţate s-ar putea să nu funcţioneze bine în modelele multivariate. Mai mult, Hair şi colaboratorii (2010) atrag atenţia că ştergerea sau modificarea cazurilor extreme poate avea un efect pervers grav : modelele multivariate vor fi mai bune din punct de vedere statistic, dar mai puţin generalizabile la populaţia pentru care facem inferenţe. Vă recomand să consultaţi aceste două lucrări, pentru că oferă exemple detaliate despre cum se identifică şi gestionează cazurile extreme.

Deoarece depăşeşte scopul acestei lucrări, nu mai insistăm asupra aspectelor teoretice ale acestei teme, aşadar vom discuta în continuare doar despre partea operaţională a identificării rapide a cazurilor extreme. Am discutat despre cazurile extreme univariate. În figura 7.3 am observat că o persoană care are id-ul 869 este considerată caz extrem. În SPSS, putem consulta imediat răspunsurile oferite de acest respondent la diferite variabile din chestionar. O variantă presupune să

mergem în **Data View**, unde dăm click în prima celulă din dreptul variabilei V3 (cea care conţine id-urile respondenţilor), apăsăm pe iconiţa şi tastăm în câmpul **Find** 869 (figura 7.4). Apoi apăsăm butonul **Find Next**, comandă care ne va duce la celula din V3 care conţine numărul 869.

Figura 7.4. Find



Putem selecta rândul care conţine acest respondent dând click pe numărul rândului, după care vom naviga folosind bara orizontală de scroll. Mai simplu ar fi însă ca ID-ul să rămână vizibil, schimbând doar poziţia celorlalte variabile. Pentru aceasta, mergem în meniul **Window > Split** (figura 7.5).

Figura 7.5. Meniul Window > Split

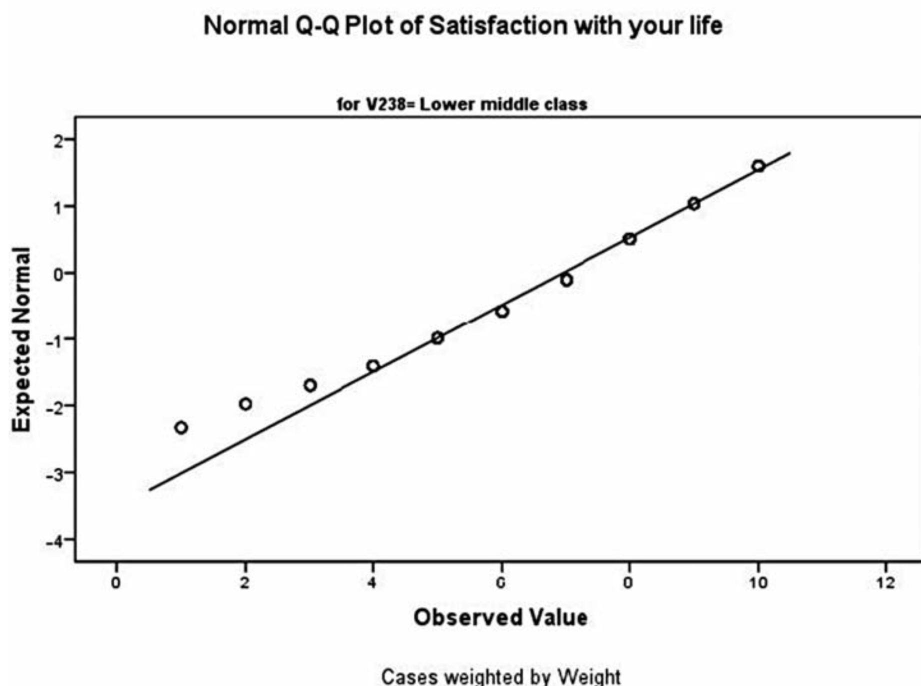
V3	V4	V6	V7	V8	V9
869	1	1	4	3	2
870	1	2	3	2	3
871	1	1	3	1	1
872	1	3	2	2	3

Fereastra se împarte în două sau patru secţiuni, în funcţie de varianta de SPSS cu care lucraţi, fiecare fiind navigabilă de sine stătător. Acum, de exemplu, ţinând constantă poziţia lui V3, în cadranul din stânga sus, putem naviga orizontal în cadranul din dreapta pentru a vedea ce valori ia respondentul cu ID-ul 869 la alte variabile. Există posibilitatea ca acea valoare extremă să fie doar o eroare de introducere a datelor. Adică operatorul de introducere, în loc să tasteze valoarea 5, a tastat valoarea 1. La variabilele subiective (valori, atitudini, evaluări etc.) este greu să ne dăm seama de aceste lucruri, dar la venituri sau proprietăţi s-ar putea să fie mai uşor. Cert este că, înainte de a lua o decizie de transformare sau ştergere, trebuie să verificăm chestionarele în original (dacă putem).

Am insistat asupra cazurilor extreme pentru că, de multe ori, asumpţia de normalitate a distribuţiei este încălcată pentru că acestea există.

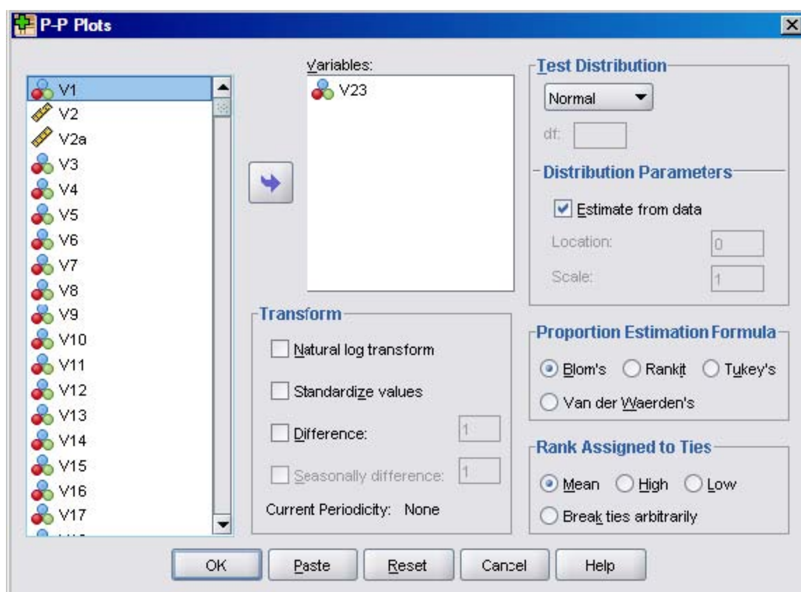
Histograma şi box-plot-ul sunt rapid de construit şi uşor de interpretat. Un alt grafic, creat special pentru evaluarea acestei asumpţii, şi care nu are dezavantajale celorlalte două, este **normal probability plot**. În figura 7.5 este prezentat un grafic similar, **normal q-q plot**, pentru satisfacţia cu viaţa în cadrul categoriei „partea de jos a clasei mijlocii". Distribuţia normală este reprezentată prin linia diagonală, iar distribuţia datelor din eşantion pentru satisfacţia cu viaţa este dată de succesiunea cerculeţelor. Dacă variabila are o distribuţie normală, atunci cerculeţele ar trebui să cadă aproximativ pe linie.

Figura 7.6. Normal probability plot



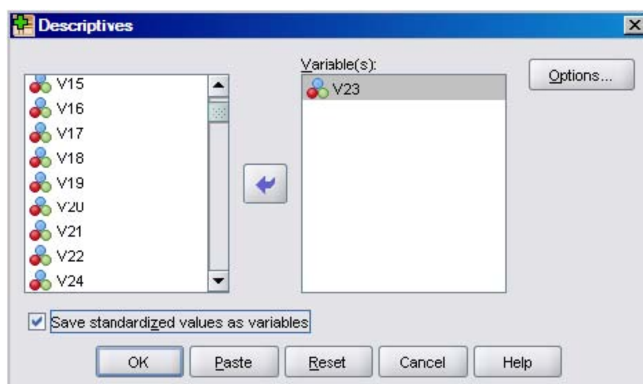
Acest grafic se obţine din meniul **Analyze > Descriptive statistics > Explore > Plots = Normality plots with tests** (Figura 7.2b). Dacă dorim să vizualizăm graficele **normal p-p**, nu pe cele **normal q-q**, atunci putem folosi alt meniu pentru a le obţine : **Analyze > Descriptive statistics > P-P Plots** sau **Q-Q Plots** (Figura 7.7). Dacă dorim să realizăm aceste grafice pentru diferite categorii ale altei variabile, aşa cum am făcut cu autopoziţionarea în ierarhia socială, atunci trebuie mai întâi să separăm (**split file**) baza de date după această variabilă sau să activăm anumite filtre (**select cases**).

Figura 7.7. Normal P-P Plots



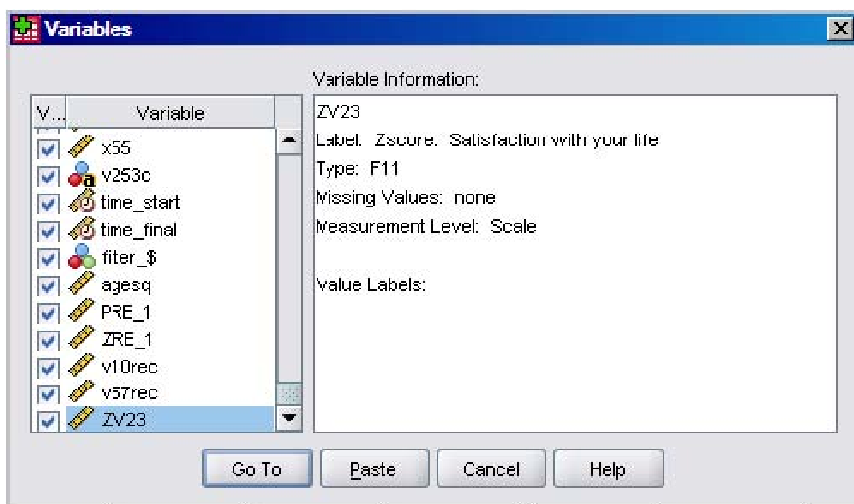
Graficele ne ajută, dar, deseori, avem nevoie şi de indicatori specifici creaţi pentru aceleaşi scopuri. Aş vrea să ne întoarcem puţin la cazurile extreme. O metodă prin care verificăm dacă o valoare este extremă constă în transformarea acelei valori în scor z. Cazurile care au scoruri z la variabila explorată cu valori mai mari decât aproximativ 3, ignorând semnul, sunt potenţiali outlieri. Trebuie să ne reamintim aici de regula empirică aplicabilă distribuţiilor aproximativ normale (Agresti şi Finlay, 2008). Scorurile z pot fi calculate din meniul **Analyze > Descriptive statistics > Descriptives > Save standardized values as variables** (figura 7.8). Opţiunea aceasta creează o variabilă nouă.

Figura 7.8. Descriptives, scoruri z



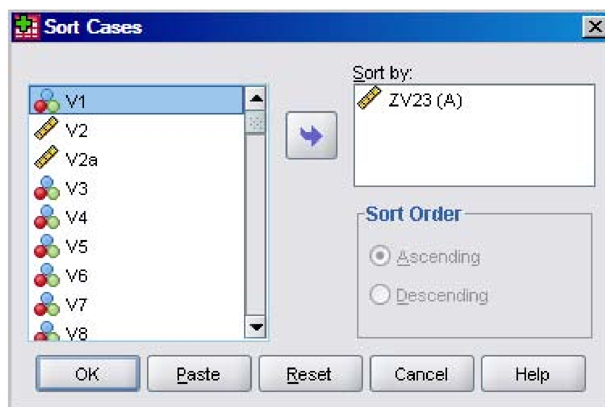
Pentru a vedea rezultatul, mergem în **Data View**, la variabila nou-creată. Aceasta va fi la sfârşitul bazei de date. Pentru a o găsi rapid, mergem în meniul **Utilities > Variables**. Se va deschide fereastra din figura 7.9.

Figura 7.9. Find variables



Selectăm orice variabilă din stânga şi tastăm rapid litera z, deoarece variabila nou-creată va fi denumită automat de SPSS care va pune această literă ca prim caracter. Dacă avem mai multe variabile standardizate, atunci tastăm rapid zv23. Apoi apăsăm **Go To** şi ne va duce la variabila dorită. Ne interesează valorile extreme, mai mari decât 3, în valoare absolută. Pentru a inspecta vizual uşor, sortăm baza de date. Mergem în meniul **Data > Sort Cases**, introducem variabila ZV23 în câmpul **Sort by** şi, la **Sort Order**, bifăm **Ascending** (figura 7.10).

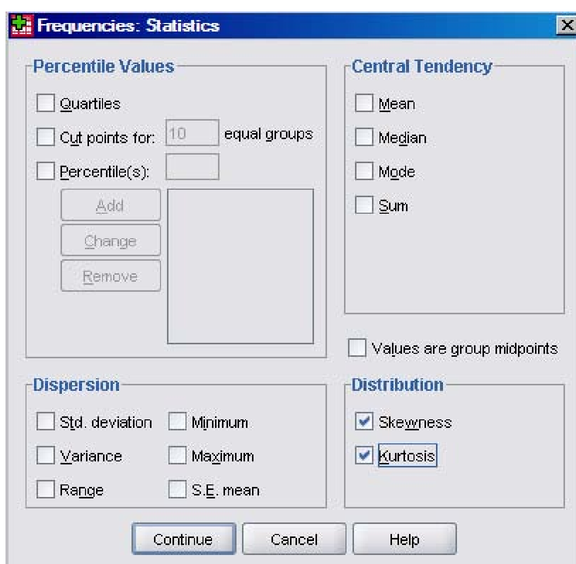
Figura 7.10. Sort Cases

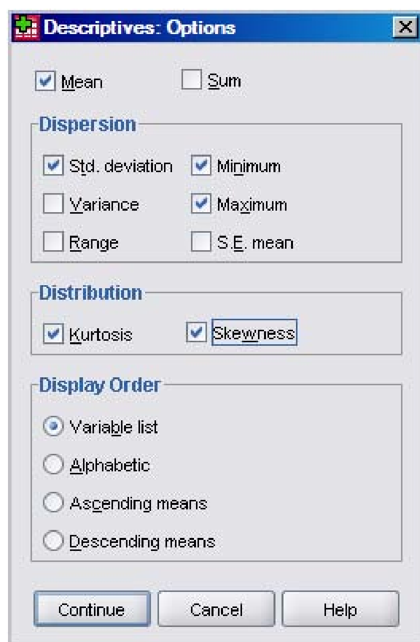


Apoi mergem în **Data View** şi inspectăm valorile. Cea mai mică este –2.37. Repetăm procedura, dar sortând **Descendindg**. Cea mai mare este 1.38. După criteriul z > 3 (în valoare absolută), nu avem cazuri extreme. Observăm cum cercetătorul trebuie să îşi folosească raţiunea atunci când decide ce este şi ce nu este caz extrem. Nu lăsăm programul să ia decizii pentru noi. Ca regulă de lucru, verificarea prin aplicarea mai multor tehnici cu acelaşi obiectiv este esenţială.

Doi dintre cei mai utilizaţi indicatori statistici pentru verificarea normalităţii distribuţiei sunt **skewness** (alungire) şi **kurtosis** (aplatizare). În SPSS, aceşti indicatori sunt centraţi în jurul valorii 0 ce reprezintă distribuţia normală. Când distribuţia este alungită la dreapta, indicatorul de skewness are valoare pozitivă. Când distribuţia este alungită la stânga, indicatorul de skewness are valoare negativă. Când observaţiile sunt grupate strâns în jurul mediei, indicatorul de kurtosis are valoare pozitivă. Când observaţiile sunt dispersate în jurul mediei, indicatorul de kurtosis are valoare negativă. Însă aceşti indicatori, în eşantioanele cu volum mare, pot să arate abateri de la normalitate, chiar şi atunci când acestea sunt mici (Tabachnick şi Fidell, 2007). De aceea, interpretarea lor trebuie combinată cu ceea ce ne oferă graficele discutate. Aceşti indicatori pot fi obţinuţi din mai multe meniuri : **Frequencies**, **Descriptives** sau **Explore**. Dacă ne interesează valorile pentru o variabilă în cadrul întregului eşantion, atunci putem să o alegem pe oricare dintre ele. Dacă vrem însă să aflăm aceste valori pentru diferite categorii ale altei variabile, ar fi mai util să folosim **Explore**. Motivul este simplu : pentru **Frequencies** şi **Descriptives** ar trebui, în prealabil, să separăm (**split file**) baza de date. Am face o operaţie în plus. Figura 7.11 prezintă opţiunile din **Frequencies** şi **Descriptives**. Meniul **Explore** le calculează implicit.

Figura 7.11. Meniurile Frequencies şi Descriptives : calcularea skewness şi kurtosis





Aici, indicatorii **skewness** şi **kurtosis** arată uşoare abateri, dar nimic grav : valori absolute mai mici decât 1 la skewness pentru toate poziţiile sociale şi doar o valoare de aproximativ 1.5 la kurtosis pentru partea de sus a clasei mijlocii (tabelul 7.2).

Putem combina valorile de la skewness şi kurtosis cu cele de la medie, mediană şi medie. Dacă media şi mediana au valori apropiate, este mai probabil să nu avem cazuri extreme. Abaterea standard ne va ajuta să înţelegem, de asemenea, cât de omogene sunt grupurile. Să ţinem minte însă că am considerat scala de 10 puncte ca fiind metrică. Rezultatele pot fi influenţate de acest lucru. De asemenea, să ţinem minte că nu toate variabilele au o distribuţie naturală normală, deci să nu căutăm normalitate acolo unde este greu de găsit.

Tabelul 7.2. Skewness şi kurtosis. Calcule efectuate în meniul Explore. Tabele obţinute prin pivotare

Descriptives			
Statistics= Skewness			
	V238 Social class (subjective)	Statistic	Std. Error
V23 Satisfaction with your life	1 Upper class	-.532	.553
	2 Upper middle class	-.927	.141
	3 Lower middle class	-.641	.117
	4 Working class	-.477	.106
	5 Lower class	.030	.192

Descriptives			
Statistics= Kurtosis			
	V238 Social class (subjective)	Statistic	Std. Error
V23 Satisfaction with your life	1 Upper class	-1.038	1.069
	2 Upper middle class	1.472	.281
	3 Lower middle class	.446	.233
	4 Working class	-.452	.211
	5 Lower class	-.945	.381

7.2. Relaţia liniară dintre două variabile

Această asumpţie priveşte relaţiile dintre două variabile. Asumpţia discutată anterior, chiar dacă am făcut trimiteri şi către normalitatea multivariată, a fost dezvoltată într-un cadru univariat.

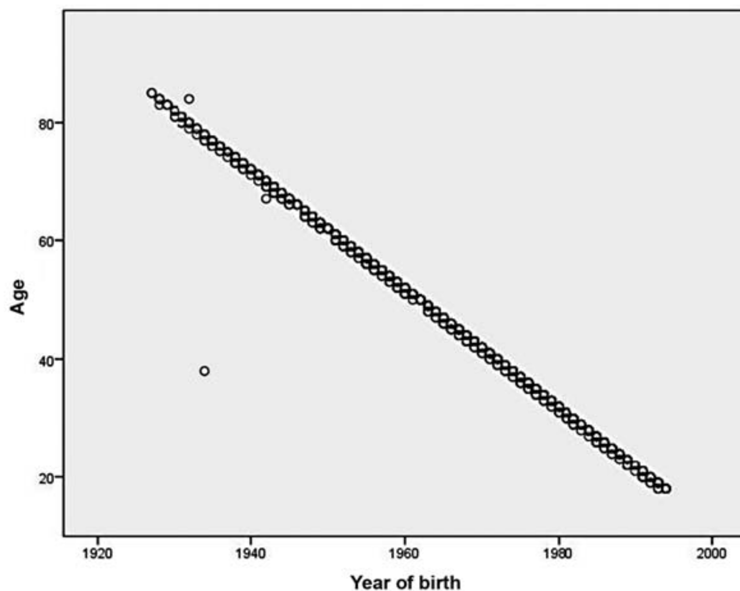
Analiza de corelaţie, care calculează coeficientul de corelaţie Pearson r, este frecvent utilizată în cercetarea socială. Aceasta ne arată dacă, între două variabile metrice, există o relaţie : la o anumită valoare a variabilei X, variabila Y ia o anumită valoare. Mai general, pentru o mulţime de persoane, dacă valorile variabilei X cresc sau scad, atunci cresc sau scad şi valorile variabilei Y (relaţie direct proporţională), sau dacă valorile variabilei X cresc sau scad, atunci scad sau cresc valorile variabilei Y (relaţie invers proporţională). Coeficientul de corelaţie Pearson r ia valori în intervalul $[-1, 1]$: când $r = -1$, relaţia este perfect negativă, când $r = 0$, între X şi Y nu există o relaţie, iar când $r = 1$, relaţia este perfect pozitivă. Analiza este simplu de interpretat. Acest lucru o face şi atractivă, probabil. Totuşi, este foarte uşor să greşim atunci când interpretăm coeficientul de corelaţie Pearson r dacă nu am verificat două asumpţii esenţiale ale acestei analize. Prima asumpţie se referă la prezenţa cazurilor extreme (outlieri). A doua asumpţie se referă la existenţa unei relaţii liniare între cele două variabile, adică o relaţie care poate fi aproximată printr-o dreaptă.

Să luăm relaţia dintre anul naşterii şi vârstă. În baza de date WVS 2012, avem ambele variabile. Vârsta este egală cu anul în care a fost aplicat chestionarul în România, 2012, minus anul naşterii respondentului. Aşadar ar trebui să avem o relaţie perfect liniară negativă : când anul naşterii creşte ca valoare, adică este mai apropiat de zilele noastre, vârsta va scădea. Ambele variabile sunt metrice, valorile pe care le pot lua cele două putând fi folosite în calcule aritmetice. Evident, acesta este un exemplu didactic care ne permite să vizualizăm o relaţie. În practică, nu ar aduce o contribuţie prea mare ştiinţei investigarea relaţiei dintre acestea. Putem vedea dacă relaţia este liniară folosind graficul **scatterplot** sau,

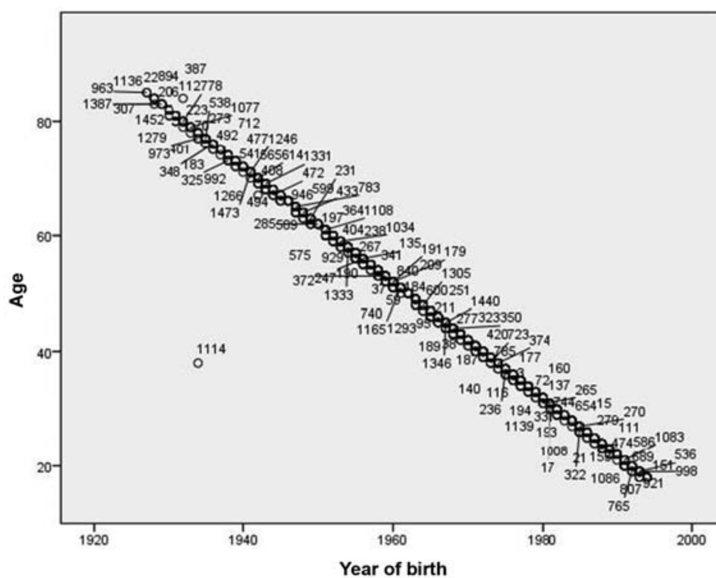
în limba română, „nor de puncte”. În figura 7.12 este prezentat acest grafic realizat pentru cele două variabile.

Figura 7.12. Scatterplot (nor de puncte) care arată o relaţie perfect liniară

(a)



(b)

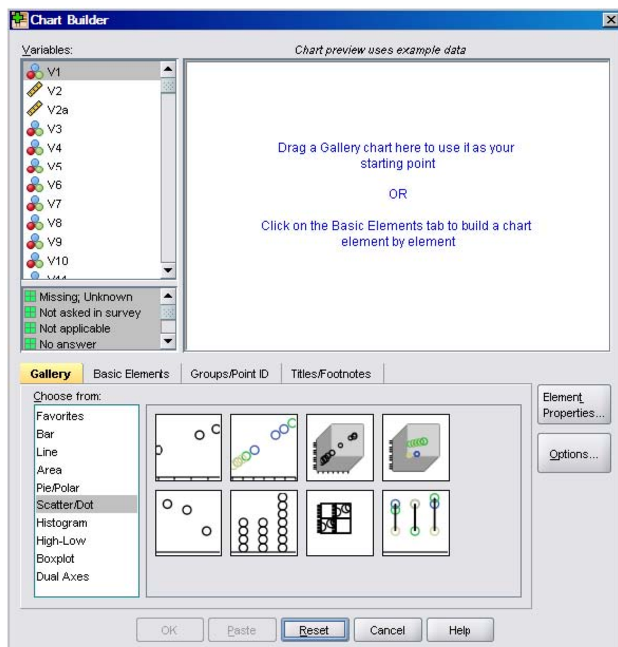


Acest grafic a fost obţinut mergând în meniul **Graphs > Chart Builder** (figura 7.13). Puteţi lucra şi cu **Graphs > Legacy Dialogs** sau **Graphs > Graphboard Template Chooser**. Rezultatele vor fi, în principiu, aceleaşi.

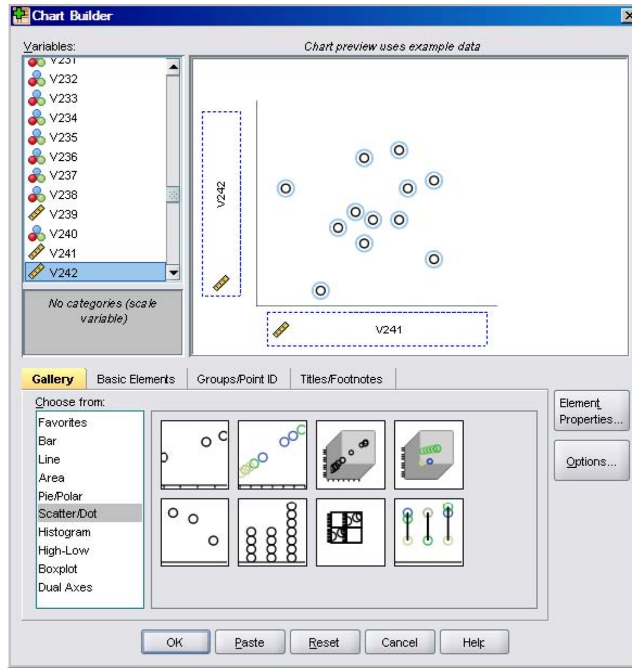
Aşa cum ne-am obişnuit, în stânga, în secţiunea **Variables** sunt toate variabilele din baza de date. De aici vom selecta, pe rând, cele două variabile : anul naşterii (V241) şi vârsta în ani împliniţi (V242). Imediat sub această fereastră, SPSS afişează valorile variabilei selectate. În partea de jos a ferestrei există patru taburi : **Gallery**, **Basic Elements**, **Groups/Point ID** şi **Titles/Footnotes**. În **Gallery** sunt graficele dintre care îl vom alege pe cel care ne interesează. Aici ne interesează scatterplotul, de aceea dăm click pe **Choose from : Scatter/Dot** (figura 7.13a). În partea dreaptă s-au activat opt tipuri de grafice (de la stânga la dreapta) : **simple scatter**, **grouped scatter**, **simple 3-d scatter**, **grouped 3-d scatter**, **summary data plot**, **simple dot plot**, **scatterplot matrix** şi **drop-line**. Noi vom utiliza graficul **simple scatter**. Mergem cu cursorul pe el şi dăm dublu click. Se va activa, în partea centrală a imaginii, structura graficului în care trebuie să introducem informaţia necesară (figura 7.13b). Pe axa X vom pune variabila pe care o considerăm explicativă. Aici nu are prea mult sens această delimitare între variabilă explicativă (independentă) şi variabilă explicată (dependentă). Dar, de dragul prezentării, vom pune anul naşterii (V241) pe axa X, pornind de la ideea că vârsta este derivată din ea. Selectăm V241 şi, prin drag-and-drop, o aducem pe axa X. Procedăm similar cu vârsta, V242, dar pe aceasta o ducem pe axa Y.

Figura 7.13. Meniul Graphs > Chart Builder

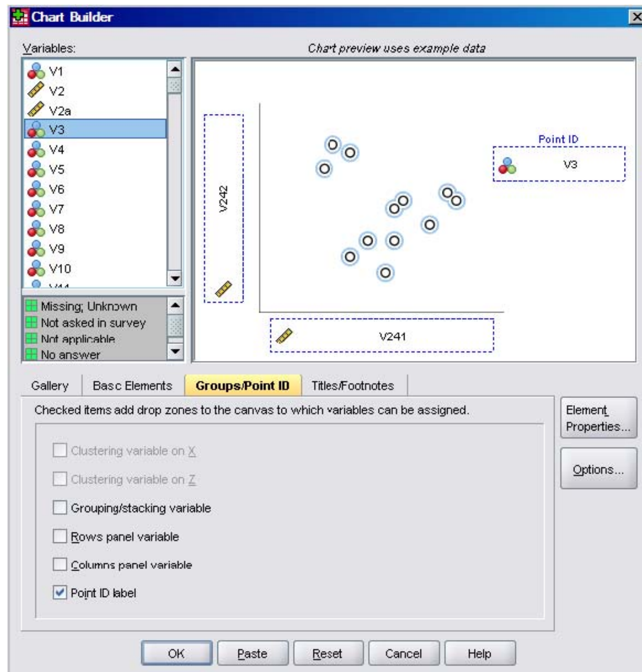
(a)



(b)


Chart Builder
Variables:
Chart preview uses example data
V232
V233
V234
V235
V236
V237
V238
V239
V240
V241
V242
No categories (scale variable)
V242
V241
Gallery
Basic Elements
Groups/Point ID
Titles/Footnotes
Choose from:
Favorites
Bar
Line
Area
Pie/Polar
Scatter/Dot
Histogram
High-Low
Boxplot
Dual Axes
Element Properties...
Options...
OK
Paste
Reset
Cancel
Help

(c)


Chart Builder
Variables:
Chart preview uses example data
V1
V2
V2a
V3
V4
V5
V6
V7
V8
V9
V10
Missing; Unknown
Not asked in survey
Not applicable
No answer
V242
Point ID
V3
V241
Gallery
Basc Elements
Groups/Point ID
Titles/Footnotes
Checked items add drop zones to the canvas to which variables can be assigned.
Clustering variable on X
Clustering variable on Z
Grouping/stacking variable
Rows panel variable
Columns panel variable
Point ID label
Element Properties...
Options...
OK
Paste
Reset
Cancel
Help

Dacă apăsăm **OK**, obţinem graficul în care observăm că există un punct ce se abate de la relaţia aşteptată. Pentru a-l putea identifica uşor, ar fi util să ştim care este id-ul său unic în baza de date. Nu uitaţi că, într-o bază de date, toate unităţile introduse trebuie să aibă un id unic. Pentru etichetare trebuie să revenim în meniul **Graphs > Chart Builder** şi să utilizăm tabul **Groups/Point ID** (figura 7.13c). Selectăm **Point ID label**. Observăm cum, în fereastra **Chart preview**, a apărut o nouă căsuţă intitulată **Point label variable ?** . Aici trebuie să introducem variabila care conţine id-urile unice ale respondenţilor. În baza de date WVS 2012, aceasta este V3. O selectăm şi, prin drag-and-drop, o aducem în căsuţa activată. Dacă apăsăm **OK**, va rezulta graficul din figura 7.12b. Aşadar, persoana care se abate de la relaţia aşteptată are id-ul unic în baza de date 1114. Pentru a vedea ce valori are această persoană la anul naşterii (V241) şi vârstă (V242), avem mai multe posibilităţi. Am putea merge în **Data View**, meniul **Window > Split**. Apoi, am putea căuta valoarea 1114 la V3 fie folosind bara verticală de navigaţie (scroll), fie folosind procedeul **Find**. Observăm o inadvertenţă : anul naşterii pentru persoana cu id = 1114 este 1934. În aceste condiţii, ne-am aştepta ca vârsta să fie egală cu 78 de ani.

Figura 7.14. Window > Split : consultarea vizuală în Data View a unor inadvertenţe în date

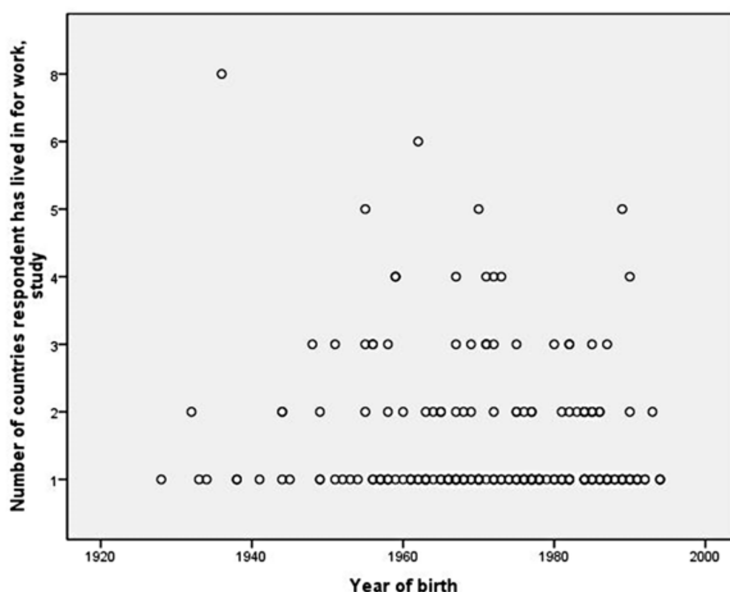
V241	V242		V3
1934	38		1114
1955	57		1115
1993	19		1120

Dar aici este egală cu 38 de ani. Dacă nu ar exista astfel de erori, corelaţia Pearson dintre anul naşterii şi vârstă ar fi egală cu –1. Acum însă este redusă la –0.99. Aici nu simţim foarte clar efectul cazurilor extreme pentru că, practic, realizăm o corelaţie a unei variabile cu ea însăşi. Însă, atunci când corelăm două variabile care măsoară lucruri diferite, efectul cazului extrem ar putea fi dramatic. O a doua metodă prin care putem vedea ce valori ia cazul cu id-ul 1114 la V241 şi V242 presupune următoarele : activăm un filtru care respectă condiţia V3 = 1114 şi apoi realizăm câte un tabel de frecvenţă pentru fiecare dintre cele două variabile. Prima metodă este preferată de începătorii care vor să vadă datele. A doua metodă ar trebui însă să fie cea pentru care optăm deoarece ne permite să salvăm sintaxele celor două acţiuni păstrând astfel şi jurnalul activităţii de analiză.

Vizualizarea relaţiilor liniare prin utilizarea scatterploturilor este posibilă doar când ambele variabile sunt cu adevărat metrice şi au valori multe. Observăm în figura 7.15 un scatterplot cu anul naşterii pe axa X şi numărul de ţări în care

respondentul a muncit sau studiat. Deşi ambele sunt metrice, numărul de ţări are o distribuţie în formă de J, adică majoritatea persoanelor indică un număr mic de ţări. Este greu să apreciem, din acest grafic, ce fel de relaţie este între cele două variabile.

Figura 7.15. Limitele scatterplotului : când o variabilă are puţine valori



Jaccard şi Jacoby (2010) oferă o explicaţie frumoasă şi uşor de înţeles a funcţiei liniare. Totuşi, acest subiect capătă şi mai mult sens dacă se trece într-un cadru multivariat. Regresia liniară multiplă are mai multe asumpţii, iar înţelegerea acestora şi a metodelor lor de testare va face mult mai clar modul în care putem depista relaţii nonliniare între variabilele care ne interesează (Berry, 1993).

7.3. Soluţii la încălcarea asumpţiei de normalitate a distribuţiei

Deseori, asumpţia normalităţii este încălcată datorită prezenţei cazurilor extreme. Uneori, de exemplu, când sunt foarte puţine, putem şterge cazurile extreme, rezolvând astfel şi problema normalităţii. Alteori însă lucrurile nu sunt atât de simple.

O soluţie des utilizată de cercetătorii experimentaţi, atunci când asumpţia de normalitate este încălcată, constă în transformarea matematică a uneia dintre

variabile sau chiar a mai multora. Hair *et al.* (2010), Tabachnick şi Fidell (2007) sau Field (2007) sunt doar câteva dintre lucrările în care ne sunt prezentate alternativele pe care le avem la îndemână. Trebuie însă să reţinem că aceste transformări nu sunt soluţii minune. De exemplu, interpretarea modelelor în care sunt folosite variabile transformate este mai dificilă decât atunci când utilizăm unităţile de măsură originale.

Pentru că depăşeşte scopul acestei lucrări, vom reda doar câteva dintre transformările uzuale, aşa cum sunt recomandate de autorii citaţi :

- distribuţie alungită la dreapta, distanţa dintre valorile minime şi maxime „normale" este mică : logaritmăm variabila ;
- distribuţie alungită la dreapta, distanţa dintre valorile minime şi maxime „normale" este ceva mai mare : radical din variabilă ;
- distribuţie alungită la dreapta, cu formă care aproximează litera J întoarsă (censored) : calculăm raportul dintre 1 şi variabilă (1/variabilă).

O lucrare foarte utilă pentru cei care vor să pătrundă tainele acestor probleme matematice îi aparţine lui John Fox (2009).

7.4. Exerciţii

Pentru aceste exerciţii utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultată(e) în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a *Grupului Românesc pentru Studiul Valorilor Sociale* (http : //www.romanianvalues.ro).

1. Verificaţi asumpţiile pentru toate exerciţiile de la capitolul 6.
2. Propuneţi soluţii de îmbunătăţire a situaţiei acolo unde este cazul.

8. Corelaţia şi regresia liniară multiplă

Care este relaţia dintre veniturile unei persoane şi numărul anilor de educaţie formală absolviţi? Are sens investiţia de timp şi resurse în educaţie? Cresc veniturile odată cu numărul anilor de educaţie formală absolviţi? Notele primite la testul-grilă cresc odată cu numărul de cursuri şi seminarii la care studenţii participă? Sau, mai degrabă, notele tind să fie mai mari atunci când studenţii petrec mai multe ore studiind individual? Satisfacţia cu viaţa este mai ridicată atunci când persoanele consideră că au control asupra propriei vieţi? Acesta este un tip de întrebări pe care ni le punem frecvent în procesul de cercetare. De fapt, abstractizând, ne întrebăm dacă între două variabile există o corelaţie. Termenul corelaţie este nou, dar ideea nu, aceasta devenind familiară deja de la asocierea testată prin tabele de contingenţă şi chi square (hi pătrat).

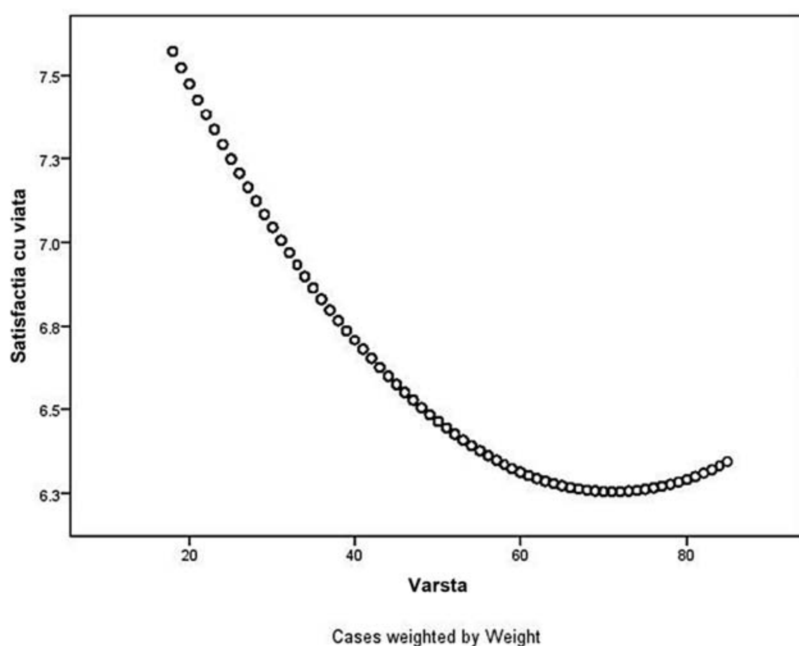
Acesta este un pas intermediar către analiza multivariată. De fapt, viaţa socială este complexă şi nu poate fi explicată adecvat folosind analize bivariate. Trebuie să învăţăm cum putem utiliza, simultan, mai mulţi predictori pentru aceeaşi variabilă dependentă. Veniturile unei persoane depind atât de educaţia formală acumulată, cât şi de capitalul cultural moştenit de la părinţi, generaţia din care face parte, sexul acesteia, vârsta, sistemul de valori la care aderă, tipul localităţii în care locuieşte etc. Notele primite la testul-grilă depind de prezenţa fizică la ore deoarece studentul poate nota idei şi exemple oferite spontan de profesor, poate pune întrebări prin care să îşi clarifice concepte şi moduri de lucru, dar şi pentru că poate discuta în pauze cu colegii despre ceea ce au audiat în timpul întâlnirii. De asemenea, contează şi studiul individual. Dar toate aceste relaţii se pot modifica dacă luăm în considerare, de exemplu, gradul de extraversiune al studenţilor. O persoană care consideră că are control asupra propriei vieţi ar putea fi mai mulţumită cu viaţa sa. Dar mulţumirea depinde şi de starea de sănătate, de starea civilă, situaţia financiară etc. Aşadar, o explicaţie cere considerarea simultană a mai multor factori care determină variaţia fenomenului care ne interesează. Una dintre cele mai utilizate analize multivariate este regresia liniară multiplă.

În prima parte a capitolului vom discuta despre corelaţia liniară necesară pentru înţelegerea regresiei liniare multiple. Apoi, vom prezenta, într-o manieră nontehnică, principiile regresiei liniare multiple şi aplicabilitatea acesteia.

8.1. Corelaţia liniară

Numele analizei bivariate discutate aici implică asumpţia că, între cele două variabile pentru care calculăm coeficientul de corelaţie, există o relaţie liniară, adică o relaţie care poate fi reprezentată printr-o dreaptă. Relaţiile sunt de tipul : (a) X1 creşte, X2 creşte ; (b) X1 creşte, X2 scade ; (c) X1 scade, X2 creşte ; (d) X1 scade, X2 scade. Nu există puncte de inflexiune. Atunci când există puncte de inflexiune, relaţia nu mai este liniară. De exemplu, relaţia dintre vârstă şi satisfacţia cu viaţa nu este liniară. Pentru a înţelege mai bine, să privim figura 8.1.

Figura 8.1. Relaţia nonliniară dintre vârstă şi satisfacţia cu viaţa



Satisfacţia cu viaţa are valoare maximă începând cu 18 ani (vârsta minimă în eşantionul WVS 2012 pentru România). Aceasta descreşte constant, dar, la o anumită vârstă, pare să revină pe un trend ascendent. Acel punct de inflexiune arată o relaţie nonliniară între cele două variabile.

O altă asumpţie a corelaţiei liniare este că ambele variabile sunt cantitative, adică, în termenii nivelurilor de măsurare, de interval sau raport. În ştiinţele sociale, frecvent, scalele simple tip Likert (o întrebare cu minim 4 variante de răspuns de tipul acord/dezacord), dar şi scorurile derivate din cele compuse sunt considerate de interval, deci cantitative. De exemplu, satisfacţia cu viaţa măsurată

pe o scală de la 1 la 10 este utilizată deseori în analize în acest mod. Nu are sens calculul coeficientului de corelaţie pentru variabilele categoriale, adică nominale şi ordinale de tipul categorii ordonate.

Corelaţia nu înseamnă cauzalitate. Analiza de corelaţie ne arată doar că două variabile variază împreună, felul relaţiei (direct sau invers proporţională) şi cât de puternică este aceasta. Însă nu putem spune cu certitudine că X1 o determină pe X2 sau invers. Pentru o interpretare în termeni cauzali, cercetătorul trebuie să respecte o serie de principii chiar în designul cercetării, cum ar fi opţiunea pentru experiment sau anchetă prin chestionar. În practică, cercetătorul nu gândeşte în termenii X1 şi X2, ci în termenii X şi Y, adică o variabilă independentă şi una dependentă. Adecvarea şi consistenţa interpretării ţine de corectitudinea logicii cercetătorului. Un domeniu în care erorile de interpretare ale analizei de corelaţie sunt foarte posibile este cel al fericirii şi satisfacţiei cu viaţa. Care este, spre exemplu, relaţia dintre starea de sănătate percepută şi satisfacţia cu viaţa ? O persoană despre care se consideră că este mai sănătoasă va fi mai satisfăcută cu viaţa sau o persoană mai satisfăcută cu viaţa se va considera mai sănătoasă ? Răspunsul nu este unul simplu, ambele variante având un anumit grad de plauzibilitate. În multe studii însă, satisfacţia cu viaţa este considerată variabila dependentă, iar starea subiectivă de sănătate este considerată variabila independentă. Starea subiectivă de sănătate depinde, în mare măsură, de starea obiectivă de sănătate (prezenţa unei boli temporare, a unei boli cronice, a unui handicap etc.), deci, dacă se intervine asupra stării obiective de sănătate, se va ajusta şi starea subiectivă de sănătate şi, într-un final, satisfacţia cu viaţa, privită ca un rezultat al vieţii de calitate. Iar o viaţă de calitate cu o stare de sănătate precară este destul de greu de imaginat.

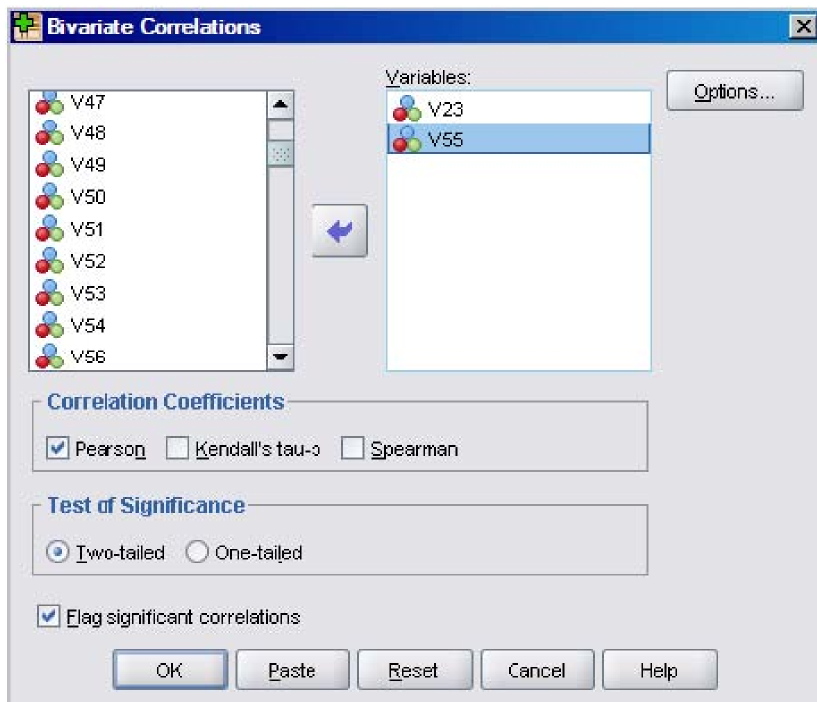
Există mai mulţi indicatori de corelaţie. Alegerea unuia depinde de caracteristicile variabilelor pe care dorim să le corelăm : variabilele corelate pot să ia multe valori, iar acestea sunt numere ; variabilele corelate conţin ranguri naturale ; variabilele corelate sunt dihotomice natural sau sunt dihotomice obţinute prin recodificarea unor variabile continue etc. Aici discutăm despre coeficientul de corelaţie Pearson r, care presupune că ambele variabile sunt cantitative continue (valorile variabilelor sunt numerice şi destul de multe). Alături de acesta, mai des întâlniţi în practică sunt coeficienţii de corelaţie Spearman, Kendall şi Gamma. Aceştia sunt specifici variabilelor ordinale, dar sunt utilizaţi şi atunci când anumite asumpţii, cum ar fi cea a distribuţiei normale bivariate, sunt încălcate. Sunt corelaţii nonparametrice, spre deosebire de Pearson, care este parametrică. O descriere foarte bună a celor mai utilizaţi coeficienţi de corelaţie a fost realizată de Chen şi Popovich (2002).

În SPSS analiza de corelaţie are un meniu dedicat : **Analyze > Correlate**. SPSS poate calcula două tipuri de corelaţii din acest meniu : corelaţie bivariată şi corelaţie parţială. În acelaşi meniu mai există opţiunea calculării unor distanţe între cazuri sau variabile. Această analiză este însă în afara intereselor noastre şi

poate fi înţeleascã mai bine în contextul analizei cluster. Meniul pentru corelaţia bivariată, **Analyze > Correlate > Bivariate**, ne permite calcularea a trei coeficienţi de corelaţie : **Pearson**, **Kendall's tau-b** şi **Spearman** (figura 18.2).

Figura 8.2. Meniul corelaţiei bivariate (Correlate > Bivariate)

(a)



(b)



În secţiunea din stânga sus sunt toate variabilele din baza de date. De aici selectăm variabilele pe care dorim să le corelăm şi, folosind săgeata dintre cele două secţiuni, le trecem în secţiunea din dreapta sus. Trebuie să introducem minim două variabile. Nu este obligatoriu să introduceţi doar două. Dacă introducem opt variabile pentru care dorim să calculăm coeficientul de corelaţie Pearson, atunci va rezulta un tabel cu opt rânduri şi opt coloane, fiecare dintre cele opt variabile fiind corelată, pe rând, cu ea însăşi şi cu toate celelalte şapte. Aşadar, indiferent de numărul de variabile pe care le introducem în analiza corelaţiei, rezultatul va fi tot bivariat. În figura 8.2 am introdus doar două variabile, v23 şi v55. Prima reflectă răspunsurile la întrebarea din WVS 2012 „Dacă luaţi în considerare toate aspectele vieţii dvs. din ultimul timp, în ce măsură sunteţi mulţumit de ea ? (utilizaţi scala de mai jos, în care 1 înseamnă «total nemulţumit» şi 10 «total mulţumit»)". A doua reflectă răspunsurile la întrebarea din aceeaşi cercetare „Unii oameni cred că au libertate totală de alegere şi de control asupra vieţii lor, iar alţi oameni cred că, indiferent ce fac, nu pot influenţa ce li se întâmplă în viaţă. Vă rugăm să folosiţi scala următoare pentru a indica câtă libertate de alegere credeţi că aveţi dvs., dând o notă de la 1 la 10, unde 1 înseamnă că «Nu am deloc», iar 10 că «Am libertate deplină»".

Următorul pas constă în alegerea coeficientului de corelaţie pe care dorim să-l calculăm. Interpretarea coeficientului de corelaţie este relativ simplă şi directă. Dacă valoarea p calculată este mai mică sau egală cu pragul teoretic de 0.05, atunci consultăm semnul coeficientului, care ne spune direcţia relaţiei, urmând ca mai apoi să interpretăm puterea relaţiei dată de valoarea absolută a coeficientului de corelaţie.

Toţi cei trei coeficienţi, Pearson, Kendall tau-b şi Spearman, variază între [–1, 1].

Interpretarea semnului se face în funcţie de semnificaţia valorilor pe care le iau cele două variabile analizate. Dacă variabilele sunt numere, atunci interpretarea este simplă. Când numărul de ore petrecute studiind individual la statistică creşte, ne aşteptăm ca notele luate la teste să crească. Semnul va fi plus. Când numărul anilor de educaţie formală creşte, ne aşteptăm ca veniturile persoanei să crească. Semnul va fi plus. Când numărul sortimentelor vândute într-un magazin este mare, ne aşteptăm ca numărul clienţilor acelui magazin să fie mare. Semnul va fi plus. Dar dacă calculăm coeficientul de corelaţie Pearson pentru două variabile măsurate fiecare pe o scală de 10 puncte, atunci trebuie să citim cu atenţie etichetele atribuite codurilor. Dacă 1 înseamnă satisfacţie scăzută cu viaţa şi 10 satisfacţie ridicată, iar 1 înseamnă absenţa controlului asupra propriei vieţi şi 10 înseamnă control total, atunci semnul va fi plus. Dacă una dintre cele două variabile ar fi codificată în alt sens, de exemplu 1 ar însemna satisfacţie ridicată cu viaţa şi 10 satisfacţie scăzută, iar la control scala s-ar păstra, atunci semnul ar fi minus.

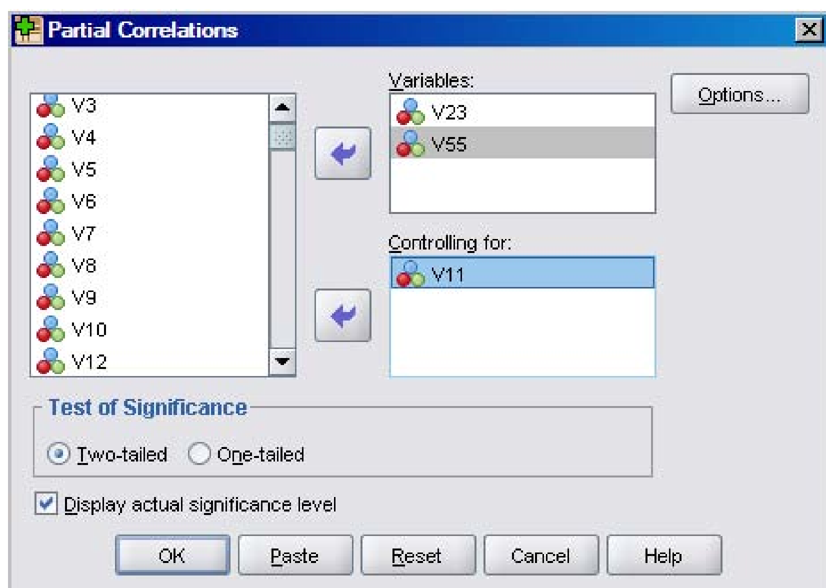
Valorile extreme indică relaţii perfecte de interdependenţă între cele două variabile. Valoarea 0 poate însemna absenţa unei relaţii de interdependenţă. Dar,

dacă ne reamintim de asumpţia liniarităţii, dar şi de cea a distribuţiei bivariate normale, un coeficient Pearson egal cu zero poate sugera şi o relaţie nonliniară, atunci când aceste asumpţii nu sunt respectate. Adică între cele două variabile există o relaţie care nu este liniară (figura 8.1), deci nu poate fi reprezentată numeric corect prin coeficientul Pearson. Nu există o regulă larg acceptată despre relaţia dintre valoarea coeficientului de corelaţie şi tăria corelaţiei. De regulă, valorile absolute mai mici de 0.3 sunt considerate corelaţii slabe spre moderate, între 0.3 şi mai mici de 0.6 sunt considerate corelaţii moderate spre puternice, iar mai mari sau egale cu 0.6 sunt considerate corelaţii puternice. Totuşi, interpretarea depinde de domeniul studiat, iar aceste valori au un caracter orientativ. Trebuie ţinut cont şi de forma distribuţiilor celor două variabile : când sunt diferite de cea normală şi/sau diferite între ele, atunci valorile maxime, –1 sau 1, sunt mai greu de atins (Carroll, 1961). De asemenea, Chen şi Popovich (2002) atrag atenţia că în eşantioanele mici, de câteva zeci de cazuri, este foarte probabil să avem coeficienţi de corelaţie cu valori mari, chiar dacă în populaţie valorile sunt mici sau corelaţia este inexistentă. În aceeaşi situaţie, trebuie să fim atenţi şi la cazurile extreme care pot afecta mărimea sau chiar direcţia coeficientului de corelaţie Pearson. Lucrarea celor doi autori detaliază toate problemele care pot influenţa rezultatul analizei de corelaţie, atunci când folosim coeficientul de corelaţie Pearson.

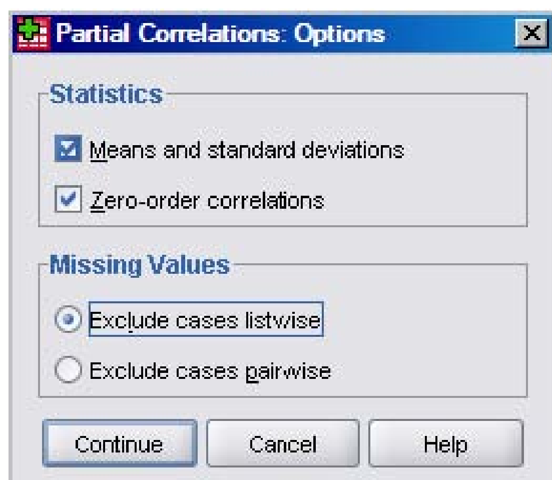
O altă problemă pe care trebuie să o avem în vedere când interpretăm un coeficient de corelaţie Pearson se referă la relaţia iluzorie dintre cele două variabile corelate. Relaţia iluzorie (*spurious*) atestă existenţa a cel puţin unei alte variabile, X3, care explică relaţia dintre X1 şi X2. Kline (2011) arată cum relaţia dintre numărul de cuvinte pe care un copil îl are în vocabular (X1) şi mărimea pe care o poartă la pantof (X2) este iluzorie, ambele fiind, de fapt, determinate de vârsta copilului (X3) folosită aici ca proxy pentru maturizarea educaţională şi fizică. Astfel de situaţii trebuie identificate teoretic de cercetător în acord cu literatura consultată sau, atunci când aceasta lipseşte, cu intuiţia proprie. Practic, poate fi testată folosind corelaţia parţială. Însă, deşi pare atrăgătoare această metodă de testare a veridicităţii relaţiei dintre două variabile, în realitate este insuficientă. Viaţa socială este mult mai complexă, fiind foarte probabil să existe mai mult de o variabilă care explică relaţia dintre cele două variabile corelate. Aceste situaţii pot fi testate mai adecvat în contextul modelelor de ecuaţii structurale (*structural equation modeling*), un subiect care depăşeşte tematica acestui volum. În SPSS, corelaţia parţială poate fi găsită în meniul **Analyze > Correlate > Partial** (figura 8.3).

Figura 8.3. Meniul corelaţiei parţiale (Correlate > Partial)

(a)



(b)



În secţiunea din stânga a ferestrei, care se deschide prin activarea meniului, se găsesc variabilele din baza de date. În secţiunea denumită **Variables,** introducem cele două variabile pe care dorim să le controlăm. Aici am introdus V23, satisfacţia cu viaţa, şi V55, controlul perceput asupra propriei vieţi. În secţiunea denumită **Controlling for** introducem variabila despre care presupunem că influenţează relaţia dintre V23 şi V55. Aici am introdus V11, care cuprinde răspunsurile la

întrebarea din WVS 2012 : „Cum aţi descrie starea dvs. de sănătate în prezent ? 1 Foarte bună, 2 Bună, 3 Nu prea bună, 4 Proastă". Presupunem că persoanele care consideră că au o stare de sănătate pozitivă apreciază că au un control ridicat asupra propriei vieţi. Dacă se întâmplă acest lucru, atunci ne aşteptăm ca relaţia dintre controlul perceput asupra propriei vieţi şi satisfacţia cu viaţa să se diminueze, adică valoarea coeficientului de corelaţie Pearson să scadă. Rughiniş (2007) dă mai multe exemple de corelaţii iluzorii, explicând care este rolul controlului variabilelor în analiză. Astfel, putem explica relaţia dintre numărul de pompieri la locul unui incendiu şi valoarea pagubelor produse dacă luăm în considerare mărimea incendiului. La fel, putem înţelege relaţia dintre numărul bisericilor dintr-un oraş şi numărul crimelor violente, ambele fiind determinate de mărimea localităţii.

În meniul corelaţiei bivariate (figura 8.2), dar şi în cel al corelaţiei parţiale (figura 8.3), mai avem câteva opţiuni pe care le putem bifa sau debifa.

Una dintre ele se referă la tipul testului de semnificaţie : **two-tailed** sau **one-tailed**. Opţiunea implicită în SPSS este two-tailed. Cercetătorul poate păstra sau modifica această opţiune în funcţie de ipoteza pe care o testează. Pentru o înţelegere a logicii testelor de semnificaţie, poate fi consultat orice manual de statistică sau lucrările dedicate acestui subiect de către Henkel (1976) şi Mohr (1990). În practică, aşa cum subliniază Mohr (1990), majoritatea ipotezelor sunt direcţionale, adică presupunem că relaţia are o anumită direcţie. Din acest motiv, ar trebui să selectăm one-tailed. Însă acelaşi autor consideră că mulţi cercetători adoptă o perspectivă mai conservatoare şi aleg opţiunea two-tailed, ca şi când natura relaţiei nu ar putea fi prezisă, folosind termenii lui Field (2009). De exemplu, dacă credem că persoanele care consideră că au control asupra propriei vieţi vor fi mai satisfăcute cu viaţa lor, atunci aş putea alege opţiunea one-tailed. Dar dacă nu ştim la ce să ne aşteptăm, alegem opţiunea two-tailed. E preferabil să alegem varianta conservatoare şi să păstrăm opţiunea implicită din SPSS.

O altă opţiune este **Flag significant correlations** în meniul corelaţiei bivariate (figura 8.2) şi **Display actual significance level** în meniul corelaţiei parţiale (figura 8.3). Ambele opţiuni au efect doar asupra modului de prezentare a tabelelor în Output. Prefer opţiunea implicită din SPSS. În tabelul 8.1 sunt prezentate rezultatele cu şi fără aceste opţiuni bifate. În cazul corelaţiei bivariate, când păstrăm bifată opţiunea implicită, în dreptul coeficientului de corelaţie sunt notate una sau mai multe steluţe (*, **), în funcţie de valoarea pe care o ia nivelul de semnificaţie. Când lucrăm, acest ajutor vizual poate fi foarte util, de aceea recomand utilizarea sa. În cazul corelaţiei parţiale, lucrurile stau invers : prin debifare sunt afişate steluţele în defavoarea valorii nivelului de semnificaţie. Tabelul afişat este mai puţin complex, dar dacă în lucrarea pe care o pregătim trebuie să raportăm chiar nivelul de semnificaţie calculat, atunci am avea nevoie să păstrăm bifată opţiunea implicită.

Tabelul 8.1. Output cu sau fără opţiunile Flag… sau Display… în meniurile corelaţiei bivariate, respectiv corelaţiei parţiale

Correlations

		V23 Satisfaction with your life	V55 How much freedom of choice and control over own life
V23 Satisfaction with your life	Pearson Correlation	1	.333**
	Sig. (2-tailed)		.000
	N	1491	1474
V55 How much freedom of choice and control over own life	Pearson Correlation	.333**	1
	Sig. (2-tailed)	.000	
	N	1474	1484

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		V23 Satisfaction with your life	V55 How much freedom of choice and control over own life
V23 Satisfaction with your life	Pearson Correlation	1	.333
	Sig. (2-tailed)		.000
	N	1491	1474
V55 How much freedom of choice and control over own life	Pearson Correlation	.333	1
	Sig. (2-tailed)	.000	
	N	1474	1484

Correlations

Control Variables			V23 Satisfaction with your life	V55 How much freedom of choice and control over own life
V11 State of health (subjective)	V23 Satisfaction with your life	Correlation	1.000	.301
		Significance (2-tailed)	.	.000
		df	0	1471
	V55 How much freedom of choice and control over own life	Correlation	.301	1.000
		Significance (2-tailed)	.000	.
		df	1471	0

Correlations				
Control Variables			V23 Satisfaction with your life	V55 How much freedom of choice and control over own life
V11 State of health (subjective)	V23 Satisfaction with your life	Correlation	1.000	.301**
	V55 How much freedom of choice and control over own life	Correlation	.301**	1.000
**. Correlation is significant at 0.01 level				

În ambele meniuri, există butonul **Options** care activează opţiunile prezentate în figurile 8.2b şi 8.3b. Cu excepţia, **Cross-product deviations and covariances** şi **Zero-order correlations**, celelalte opţiuni sunt similare. În practică, de regulă, în cazul corelaţiei bivariate, bifăm **Means and standard deviations**, iar în cazul corelaţiei parţiale bifăm această opţiune şi **Zero-order correlations**. Prima opţiune ne afişează media şi abaterea standard pentru fiecare dintre variabilele incluse în analiză (tabelul 8.2).

Tabelul 8.2. Opţiunea Means and standard deviations din meniurile corelaţiei bivariate, respectiv corelaţiei parţiale

Descriptive Statistics			
	Mean	Std. Deviation	N
V23 Satisfaction with your life	6.70	2.385	1474
V55 How much freedom of choice and control over own life	7.88	2.279	1474

Faptul că ni se oferă posibilitatea de a calcula media şi abatarea standard pentru variabilele corelate face evident, încă o dată, că această analiză solicită variabile metrice continue. În exemplul prezentat aici, am utilizat două scale simple tip Likert cu 10 variante de răspuns şi o scală simplă tip Likert cu 4 variante de răspuns. Mediile şi, implicit, abaterile standard, calculate pentru astfel de variabile au un caracter mai degrabă artificial fiind, uneori, chiar dificil de interpretat (de exemplu, când scala are o variantă de mijloc evidenţiată printr-o etichetă de tipul „nici acord, nici dezacord"). Chiar dacă în practică astfel de analize sunt acceptate convenţional, trebuie să fim conştienţi de posibilele erori pe care le putem introduce în interpretările substanţiale ale unor astfel de rezultate.

Tabelul 8.3. Opţiunea Zero-order correlations în meniul corelaţiei parţiale

Correlations					
Control Variables			V23 Satisfaction with your life	V55 How much freedom of choice and control over own life	V11 State of health (subjective)
-none-[a]	V23 Satisfaction with your life	Correlation	1.000	.333**	-.365**
	V55 How much freedom of choice and control over own life	Correlation	.333**	1.000	-.154**
	V11 State of health (subjective)	Correlation	-.365**	-.154**	1.000
V11 State of health (subjective)	V23 Satisfaction with your life	Correlation	1.000	.301**	
	V55 How much freedom of choice and control over own life	Correlation	.301**	1.000	
a. Cells contain zero-order (Pearson) correlations.					
**. Correlation is significant at 0.01 level					

Zero-order correlations (tabelul 8.3) se referă la corelaţiile bivariate dintre toate variabilele pe care le includem în analiza de corelaţie parţială. Aici avem trei variabile : v23, v55 şi v11. Corelaţia parţială cu o singură variabilă de control se numeşte first-order correlation.

În fine, ultimul lucru care ne interesează, la acest nivel, este modul de tratare a nonrăspunsurilor în analiza de corelaţie. Deşi nu am menţionat până acum, bănuiesc că a fost evident că nu putem calcula coeficientul de corelaţie Pearson sau oricare altul decât după ce am instruit SPSS să dezactiveze în analize codurile de nonrăspuns (**missing**). Pentru cele trei variabile utilizate pentru exemplificare, tabelele de frecvenţe (**Analyze > Descriptive statistics > Frequencies**) arată următoarele :

Variabila	**Volumul eşantionului**	**Nonrăspunsuri**	**Volumul valid**
V23, satisfacţia cu viaţa	1503	13 / 1%	1490
V55, controlul perceput asupra propriei vieţi	1503	20 / 1%	1483
V11, evaluarea sănătăţii proprii	1503	1 / 0.1%	1502

Nonrăspunsurile (codurile -2 – „nu răspund”, respectiv -1 – „nu ştiu”) au fost scoase din analiză în **Variable View > Missing > Discrete missing values**. După această operaţiune, au fost calculaţi coeficienţii de corelaţie. Dacă ne uităm în tabelul 8.2, observăm în dreptul celor trei variabile acelaşi total, 1.474 persoane. S-a ajuns la acelaşi total selectând **Exclude cases listwise** în secţiunea **Missing Values**. Au fost ignoraţi în analiza de corelaţie indivizii care nu au oferit un răspuns valid la cel puţin una dintre cele trei variabile analizate. Dacă selectam **Exclude cases pairwise**, atunci am fi avut totaluri diferite la variabile astfel : la corelaţia bivariată 1.490 la v23 şi 1.483 la v55. În practică, pentru a nu introduce erori de interpretare, dată fiind compoziţia diferită a grupurilor, alegem să tratăm nonrăspunsurile **listwise**.

În încheiere, să interpretăm şi coeficienţii de corelaţie bivariată, respectiv cei de corelaţie parţială. Folosim informaţiile din figura 8.4.

În ambele situaţii, fie că avem o corelaţie bivariată, fie una parţială, mai întâi consultăm valoarea p (nivelul de semnificaţie), care în SPSS este notată **Sig** sau „**Significance**”. Pragul de semnificaţie este ales a priori analizei. Pragurile acceptate sunt 0.05 şi 0.01. Dacă am ales pragul cel mai puţin restrictiv, 0.05, şi observăm că p calculat este mai mic decât această valoare, atunci putem respinge ipoteza de nul a lipsei de corelaţie (am folosit varianta **two-tailed**, non-direcţională). Pentru corelaţia dintre v23 (satisfacţia cu viaţa) şi v55 (controlul perceput asupra propriei vieţi) p este egal cu 0.000. De fapt, p nu este 0, ci o valoare cu foarte multe zecimale după virgulă. Fiind mai mică decât 0.05 putem aprecia că există o corelaţie între satisfacţie şi control.

Semnul coeficientului de corelaţie Pearson este „+”, deci am fi tentaţi să spunem că ambele variabile variază în acelaşi sens. Deoarece acestea sunt măsurate prin scale cu 10 puncte, trebuie să vedem cum sunt codificate. Aici sensul este acelaşi : codul cel mai mic înseamnă situaţia negativă (satisfacţie scăzută, respectiv lipsa controlului), iar codul cel mai mare înseamnă situaţia pozitivă (satisfacţie ridicată, respectiv prezenţa unui control ridicat). Aşadar semnul pozitiv indică o relaţie pozitivă. Ne amintim că, teoretic, corelaţia nu implică cauzalitate. În practică însă, cercetătorul atribuie unei variabile rolul de dependentă, iar celeilalte de independentă. Aici am interpreta că, atunci când sentimentul de control asupra propriei vieţi creşte, creşte şi satisfacţia cu viaţa.

În fine, trebuie să apreciem cât de puternică este corelaţia : $r = 0.33$. Folosind regulile empirice întâlnite în multe surse ştiinţifice, aceasta este o corelaţie moderată.

Interpretarea este similară pentru coeficientul de corelaţie parţială. Apare ceva în plus din punct de vedere conceptual, lucru evident în tabelul 8.3. Să ne reamintim că am controlat pentru v11, evaluarea stării de sănătate, pentru că am presupus că explică parţial relaţia dintre controlul perceput asupra vieţii şi satisfacţia cu viaţa. Dacă se întâmplă aşa, atunci ne aşteptăm ca, după ce am controlat pentru v11, corelaţia dintre v23 şi v55 să scadă. Coeficientul de corelaţie bivariată dintre v23

şi v55 este egal cu 0.33, iar după ce am controlat pentru v11 acesta scade la 0.30. Diferenţa nu este mare. Putem fi entuziaşti şi să observăm scăderea, dar trebuie să fim şi realişti văzând că diferenţa nu este mare. Probabil mai există şi alţi factori care modelează relaţia dintre v23 şi v55. Dar testarea acestei idei presupune un cadru multivariat.

8.2. Regresia liniară multiplă

Calitatea vieţii unei persoane sau, în ansamblu, a unei populaţii are două componente : una obiectivă, de stare, şi una subiectivă, de evaluare (Zamfir *et al.*, 1984). Aceste două componente presupun efectuarea unor măsurători pe mai multe dimensiuni ale vieţii. La nivel individual starea sănătăţii se poate măsura, printre altele, prin prezenţa/absenţa unei boli cronice şi/sau a unei incapacităţi fizice care împiedică persoana, într-o anumită măsură, să îşi desfăşoare activităţile într-o zi obişnuită. La nivel naţional, starea sănătăţii se poate măsura, printre altele, folosind rata de morbiditate. Elaborarea unui set comprehensiv de indicatori pentru care poate fi culeasă informaţie statistică de calitate este dificilă dată fiind complexitatea dimensiunilor vieţii umane. O încercare de sistematizare este oferită de Mărginean (2005). Această perspectivă asupra calităţii vieţii ia prea puţin în considerare persoana ca fiinţă care participă la viaţa socială. De aceea, setul de indicatori de stare este completat cu o serie de indicatori de evaluare a calităţii vieţii. În anchetele dedicate calităţii vieţii, cum ar fi Diagnoza Calităţii Vieţii ICCV sau European Quality of Life Survey, indivizii sunt rugaţi să aprecieze cât de bune sau proaste sunt, de exemplu, serviciile de sănătate publice. De asemenea, sunt rugaţi să îşi exprime gradul de mulţumire cu diferite domenii ale vieţii proprii, dar şi cu viaţa în general.

Satisfacţia cu viaţa primeşte o atenţie deosebită în studiile de calitatea vieţii pentru că reflectă analiza raţională a propriei situaţii (Diener, 1984), luând în calcul simultan valorile pentru toate criteriile relevante ale standardului subiectiv al unei vieţi bune (Veenhoven, 1996). O satisfacţie cu viaţa ridicată înseamnă o calitate a vieţii ridicată. Rămâne să identificăm care sunt factorii care sporesc satisfacţia cu viaţa.

Factorii care explică satisfacţia cu viaţa pot fi grupaţi în mai multe calupuri. Un prim calup se referă la caracteristicile individuale : gen, vârstă, educaţie, stare civilă, situaţie financiară, stare de sănătate obiectivă şi autoevaluată etc. Aceşti indicatori sunt nelipsiţi, jucând, de regulă, rolul de variabile de control. Un al doilea calup se referă la mecanismele psihologice şi psihosociale care determină un nivel mai scăzut sau mai ridicat al satisfacţiei cu viaţa, cum ar fi procesul comparaţiei sociale (Michalos, 1985 ; Easterlin *et al.*, 2010) sau cel al maximizării, întâlnit în

societăţile de consum (Schwartz, 2004). Un al treilea calup se poate referi la calitatea percepută a serviciilor publice şi condiţiilor de trai din zona în care persoana locuieşte (Gandelman, Piani şi Ferre, 2012). Fără a epuiza subiectul, aş mai nota aici setul de caracteristici materiale şi/sau culturale al unei unităţi sociale cum ar fi vecinătatea (de exemplu, cartierul sau sectorul în oraşe), judeţul, regiunea de dezvoltare sau istorică ori chiar ţara (Hagerty şi Veenhoven, 2003 ; Inglehart şi Welzel, 2005 ; Inglehart *et al.*, 2008 ; Hooghe şi Vanhoutte, 2011 ; Mikucka, 2012). Din perspectiva sociologului, abordarea multinivel este necesară pentru explicarea cât mai adecvată a variaţiei satisfacţiei cu viaţa. Multinivel înseamnă includerea simultană în analiza de regresie a caracteristicilor individuale culese prin chestionare şi a caracteristicilor unităţii sociale relevante pentru studiu disponibile, de regulă, la Institutele Naţionale de Statistică sau Eurostat ori alte organizaţii internaţionale care au ca obiect de activitate agregarea indicatorilor pe care îi furnizează primăriile, spitalele, angajatorii etc. Analiza multinivel este doar o regresie, după cum spune Bickel (2007), dar, fiind ceva mai complicată, nu va fi tratată aici. Pentru a înţelege analiza multinivel, trebuie să înţelegem analiza la nivel individual.

Am putea fi interesaţi, de exemplu, să vedem cum variază satisfacţia cu viaţa în funcţie de starea materială a persoanelor, dar şi de reprezentarea despre această stare materială. Teoretic, reprezentarea despre situaţia materială ar trebui să fie consistentă cu starea concretă. Totuşi, aşa cum arată multe studii care pornesc de la teoria comparaţiei sociale, starea şi reprezentarea pot să acţioneze ca factori independenţi asupra satisfacţiei cu viaţa. De exemplu, unei persoane îi este mai degrabă teamă să piardă un lucru dobândit decât să câştige acel lucru (*loss aversion*) (Tversky şi Kahneman, 1991). Tocmai la cei care au acumulat mai multe resurse s-ar putea să acţioneze un mecanism de insatisfacţie prin modificarea în sus a standardului de referinţă (Graham şi Pettinato, 2006).

Pentru exemplificare vom utiliza datele culese în cercetarea Diagnoza Calităţii Vieţii 2003 de către Institutul de Cercetare a Calităţii Vieţii. Pentru a menţine caracterul introductiv al volumului, vom realiza o analiză de regresie care are doar două variabile independente : venitul persoanelor active pe piaţa muncii şi autopoziţionarea pe scala sărac-bogat. Variabila dependentă este satisfacţia cu viaţa. Formulările exacte din chestionar sunt :

Cât de satisfăcut sunteţi de viaţa dvs. în general ?

0	1	2	3	4	5	6	7	8	9	10
Complet nesatisfăcut										Complet satisfăcut

Vă rugăm să menţionaţi toate veniturile gospodăriei dvs. din luna trecută, mai 2003	
	Dumneavoastră
Salariul din activitatea principală	

Salariu de la un al doilea loc de muncă	
Venituri din activităţi ca întreprinzător/patron	
Venituri din activităţi ocazionale	
Venituri din proprietăţi (profit, dobânzi, dividende, chirii)	
Venituri obţinute din vânzarea produselor agricole	

respectiv

În orice societate, unii oameni se consideră bogaţi, alţii se consideră săraci. Având în vedere numerotarea de la 1 la 10, dvs. unde vă situaţi ?

1	2	3	4	5	6	7	8	9	10
Sărac									Bogat

Regresia liniară multiplă se exprimă formal prin ecuaţia :
$Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \ldots \beta_n * X_n + e$.
Pentru noi, aceasta devine :

$$\text{Satisfacţia cu viaţa} = \alpha + \beta_1 * \text{Venitul persoanelor active} + \beta_2 * \text{Autopoziţionarea pe scala sărac-bogat} + e.$$

Rulând această analiză putem răspunde următoarelor întrebări :

- Există un efect semnificativ statistic al celor două variabile independente asupra satisfacţiei cu viaţa ?
- Dacă există, care este direcţia acestei relaţii ? În ce sens se modifică satisfacţia cu viaţa când venitul persoanelor active se modifică ? Dar când autopoziţionarea pe scala sărac-bogat se modifică ?
- Cu câte unităţi se modifică satisfacţia cu viaţa atunci când venitul persoanelor active se modifică cu o unitate ? Dar când autopoziţionarea pe scala sărac-bogat se modifică cu o unitate ?
- Ce parte din varianţa satisfacţiei cu viaţa este explicată de venitul persoanelor active şi autopoziţionarea pe scala sărac-bogat ?

În termeni substanţiali, putem afla dacă situaţia materială şi/sau reprezentările despre aceasta explică satisfacţia cu viaţa şi, în caz că da, dacă efectul situaţiei materiale se păstrează atunci când controlăm pentru reprezentarea despre aceasta. De asemenea, putem deduce dacă trebuie să mai căutăm şi alţi factori explicativi ai satisfacţiei cu viaţa pe care i-am omis din analiză. Acesta este un exemplu didactic. Dacă am scrie o lucrare ştiinţifică, atunci, cu certitudine, modelul ar trebui să fie mai complex. Am exclus, de exemplu, variabile de control esenţiale cum ar fi genul, vârsta, educaţia sau alţi predictori esenţiali ai satisfacţiei cu viaţa cum ar fi evaluarea propriei stări de sănătate, evaluarea domeniilor importante

ale vieţii ş.a. Indiferent de cât de complex este modelul nostru, trebuie, înainte de a începe analiza în SPSS, să avem un model explicativ clar specificat care ţine cont de ceea ce a fost deja demonstrat în domeniu. Una dintre cerinţele esenţiale ale acestei analize este specificarea corectă a modelului, adică includerea tuturor variantelor relevante. Evident, realitatea socială este prea complexă pentru a oferi explicaţii perfecte. Dar explicaţiile parţiale pe care le producem trebuie să fie consistente. De aceea, analiza de regresie nu se face prin „încercare şi eroare". Nu deschidem baza de date şi începem să introducem şi să scoatem variabile independente în model până când rezultă ceva care seamănă cu ceea ce credeam că ar fi trebuit să rezulte. În fond, în eşantioanele cu volume mari o să găsim relaţii semnificative statistice din pură întâmplare.

Regresia liniară multiplă este doar un tip de regresie. Probabil, este cel mai utilizat tip. Opţiunea pentru un tip de regresie ţine, printre altele, de caracteristicile variabilei dependente. Dacă variabila dependentă este cantitativ continuă, atunci putem utiliza regresia liniară multiplă. Dacă este dummy (1/0), unde codul 1 este atribuit caracteristicii care ne interesează, atunci putem utiliza regresia logistică binară. Dacă este nominală cu cel puţin trei categorii, atunci putem utiliza regresia logistică multinomială. Dacă reprezintă o numărare şi are o distribuţie în formă de J întors sau J simplu, putem utiliza regresia **count**. Dacă este ordinală, putem utiliza regresia ordinală. Lista poate continua. Literatura în această zonă este bine dezvoltată. Revenind la regresia liniară multiplă, am spus că variabila dependentă trebuie să fie cantitativă continuă. În ştiinţele sociale, cu precădere, dar fără a ne limita doar la acestea, este destul de greu să identificăm instrumente de măsurare care produc variabile care iau, teoretic, o infinitate de valori. De regulă, atunci când reuşim să măsurăm cantitativ, aceastea au un caracter discret. Una dintre cele mai întâlnite proceduri de măsurare în ştiinţele sociale este scala tip Likert. Rensis Likert este unul dintre pionierii măsurării în ştiinţele sociale, propunând o scală compusă care îi poartă numele fiind, chiar şi astăzi, foarte populară (Likert, 1932). Atunci când auzim un analist spunând „scală tip Likert" nu înseamnă în mod necesar că se referă la scala compusă. Acesta s-ar putea referi la tipul variantelor de răspuns. Forma standard este Acord/Dezacord, acestea fiind extremele unei scale de răspuns cu minim patru puncte : Acord total (4), Acord (3), Dezacord (2), Dezacord total (1). Există multe variante, cu sau fără variantă de mijloc :

Dezacord total (1), Dezacord (2), Acord (3), Acord total (4)
Dezacord total (1), Dezacord (2), Nici acord, nici dezacord (3), Acord (4), Acord total (5)
Dezacord total (1) (2) (3) (4) (5) (6) Acord total (7)
Dezacord total (1) (2) (3) (4) (5) (6) (7) (8) (9) Acord total (10)
Dezacord total (1) (2) (3) (4) (5) (6) (7) (8) (9) (10) Acord total (11)
Complet nesatisfăcut (1) (2) (3) (4) (5) (6) (7) (8) (9) Complet satisfăcut (10)

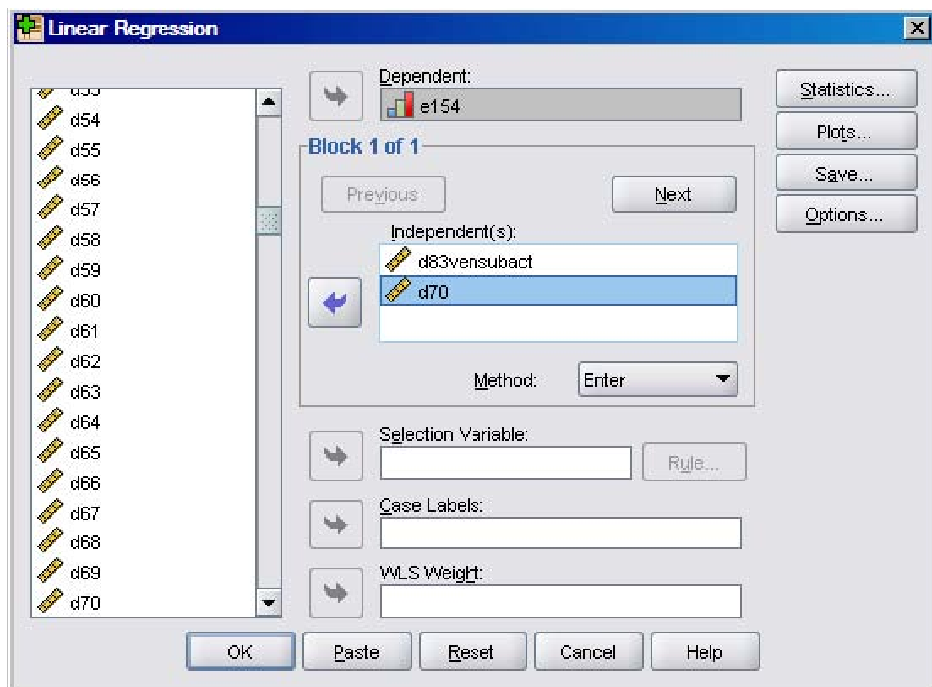
Pe lângă numărul variantelor de răspuns sau opţiunea pentru utilizarea variantei de mijloc, pot fi folosite şi etichete diferite. Acestea sunt variabile ordinale care, mai ales când au cel puţin şapte variante de răspuns, sunt considerate de interval. Fiind considerate de interval, sunt utilizate în mod curent în analizele statistice ca variabile dependente în regresia liniară multiplă. Există argumente pro şi contra (Carifio şi Perla, 2007). Referindu-se mai degrabă la variabilele numerice cu puţine valori (cantitative discrete) folosite drept variabile dependente, Berry (1993) recomandă să nu folosim variabilele cantitative discrete ca dependente în regresia liniară atunci când numărul valorilor este mai mic decât 5, iar Fox (1991), pe lângă această recomandare, fără însă a cuantifica ca Berry, consideră că mai reprezintă o problemă serioasă doar atunci când majoritatea răspunsurilor sunt concentrate pe un număr mic de valori.

O altă cerinţă esenţială a regresiei liniare este, aşa cum sugerează chiar numele analizei, ca relaţia dintre variabila dependentă şi variabilele independente să fie liniară. Dacă nu este respectată această cerinţă, atunci trebuie aplicată o formă de regresie nonliniară.

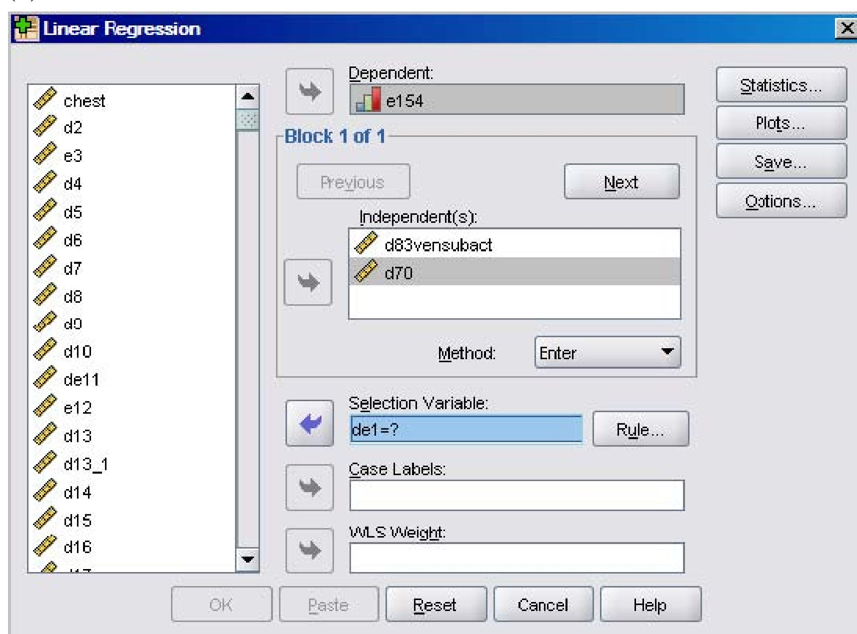
În SPSS, analiza de regresie liniară multiplă poate fi realizată din meniul **Analyze > Regression > Linear** (figura 8.4). Acesta este intuitiv.

Figura 8.4. Meniul Analyze > Regression > Linear

(a)



(b)


Linear Regression
chest
d2
e3
d4
d5
d6
d7
d8
d0
d10
de11
e12
d13
d13_1
d14
d15
d16
Dependent:
e154
Block 1 of 1
Previous
Next
Independent(s):
d83vensubact
d70
Method:
Enter
Selection Variable:
de1=?
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help


Linear Regression: Set Rule
Define Selection Rule
de1
Value:
equal to
1
Continue
Cancel
Help


Selection Variable:
de1=1
Rule...

(c)

Linear Regression: Statistics

Regression Coefficient

Estimates

Confidence intervals

Level(%): 95

Covariance matrix

Model fit

R squared change

Descriptives

Part and partial correlations

Collinearity diagnostics

Residuals

Durbin-Watson

Casewise diagnostics

Outliers outside: 3 standard deviations

All cases

Continue Cancel Help

(d)

Linear Regression: Options

Stepping Method Criteria

Use probability of F

Entry: .05 Removal: .10

Use F value

Entry: 3.84 Removal: 2.71

Include constant in equation

Missing Values

Exclude cases listwise

Exclude cases pairwise

Replace with mean

Continue Cancel Help

În secţiunea **Dependent**, introducem variabila dependentă. Aici aceasta este satisfacţia cu viaţa care, în baza de date, are numele e154.

Urmează secţiunea **Block 1 of 1 – Independent(s)**, unde introducem variabilele independente. În exemplul nostru, avem două variabile independente, venitul din ultima lună al persoanelor active pe piaţa muncii, care poartă numele d83vensubact în baza de date, respectiv autopoziţionarea pe scala sărac-bogat, care poartă numele d70 în baza de date. Cum spuneam, analiza de regresie presupune elaborarea modelului explicativ a priori utilizării programului de statistică. Putem aborda analiza în mai multe moduri. O primă variantă constă în introducerea tuturor variabilelor independente într-un singur calup sau block (am să folosesc **block** pentru a asigura corespondenţa cu programul). A doua variantă constă în gruparea, justificată teoretic, a variabilelor independente în mai multe blockuri. Am putea crea un block care conţine variabilele de control (gen, vârstă, stare civilă etc.). Apoi am putea crea un alt block care conţine informaţii despre situaţia materială a persoanei (venit, proprietăţi etc.). În fine, am putea crea un block care conţine informaţii despre cum se vede (percepe) persoana în societate din perspectiva resurselor materiale pe care le deţine (autopoziţionarea pe scala sărac-bogat, raportarea subiectivă a venitului la necesităţi etc.). Pentru că variabilele din cele trei blockuri au o utilitate proprie, surprinzând aspecte distincte de celelalte, are sens să le folosim ca atare.

Variabilele independente pot fi folosite ca atare sau pot fi grupate în scoruri compuse. Dacă teoria spune că unele variabile ar putea fi grupate sau trebuie grupate în diferite scoruri compozite sau, altfel spus, indici, atunci am fi utilizat în regresie aceşti indici. De exemplu, dacă am fi măsurat o variabilă independentă printr-o scală compusă Likert, atunci ar fi fost necesară calcularea scorului sumativ (varianta standard) sau am fi calculat media afirmaţiilor care o compune ori am fi realizat o analiză factorială exploratorie salvând scorurile factoriale pe care, ulterior, le-am fi utilizat în regresie. Există mai multe metode de calculare a indicilor, decizia aparţinând în final analistului. Acesta va trebui să pună în balanţă proprietăţile statistice ale indicelui calculat cu dificultatea de interpretare a acestuia în analiza de regresie. Scorul sumativ este mai greu de interpretat decât media variabilelor care constituie scala compusă. Indicele calculat ca medie a variabilelor variază în acelaşi interval cu cel al variantelor de răspuns, deci va fi mai uşor de înţeles. Scorul sumativ pentru o scală compusă cu 4 variabile şi 10 variante de răspuns, unde 1 = acord şi 10 = dezacord, variază între 4, dacă respondentul alege codul 1 la toate variabilele, şi 40, dacă alege codul 10 la toate variabilele. Cercetătorul trebuie să clarifice ce înseamnă scorul 13 sau scorul 33. La fel se întâmplă cu scorul factorial.

În secţiunea **Method**, avem mai multe metode, cea implicită fiind **Enter**. Aceasta este cea pe care o preferăm deoarece lasă la latitudinea cercetătorului modul în care introduce variabilele independente în analiză. Este consistentă cu elaborarea preliminară a modelului explicativ. Celelalte seamănă, mai degrabă,

cu un proces de încercare-eroare prin care analistul „caută” o relaţie semnificativă statistic.

În secţiunea **Selection Variable** putem introduce o variabilă care identifică anumite grupuri, instruind astfel SPSS-ul să ruleze analiza de regresie doar pe anumite cazuri. De exemplu, dacă doresc să realizez analiza doar pentru bărbaţi, atunci introduc variabila sex, aici **de1** (figura 8.4b). Odată introdusă variabila, se activează butonul **Rule**. De la **de1 = ?** trebuie să ajungem la **de1=1**. Adică trebuie să introducem codul care identifică grupul pentru care dorim să facem analiza de regresie. Printr-un tabel de frecvenţă am aflat că bărbaţii au codul 1 şi, deoarece dorim să rulăm regresia pentru bărbaţi, apăsăm butonul **Rule** şi introducem cifra 1 în câmpul **Value** după ce ne-am asigurat că este selectată opţiunea **equal to** în secţiunea **Define selection rule**.

În secţiunea **Case Labels** putem introduce o variabilă care identifică cazurile în mod precis în graficele pe care le realizăm odată cu celelalte calcule specifice analizei de regresie. De exemplu, am putea introduce identificatorul unic pentru fiecare respondent care, în această bază de date, se numeşte chest.

În secţiunea **WLS Weight** putem introduce o variabilă specială care ne permite rularea unui regresii liniare ajustate, utilă atunci când este încălcată asumpţia homoscedasticităţii (homoskedasticity) (Lewis-Beck, 1980).

Meniul are o serie de butoane : **Statistics**, **Plots**, **Save** şi **Options**. Vom prezenta în continuare unele dintre cele mai importante şi utile opţiuni pe care le putem alege.

Butonul **Statistics** (figura 8.4c) conţine informaţiile esenţiale pentru care alegem să rulăm această analiză. Implicit sunt selectate, în secţiunea **Regression Coefficient**, **Estimates** şi, alături, **Model fit**. **Estimates** ne va afişa coeficienţii de regresie nestandardizaţi şi coeficienţii de regresie standardizaţi. Coeficienţii de regresie nestandardizaţi (acei β din ecuaţia de regresie) ne arată cu cât se modifică variabila dependentă atunci când variabila independentă corespunzătoare se modifică cu o unitate. Aceştia pot avea semnul minus sau plus, în funcţie de relaţia dintre X şi Y, dar şi de modul în care sunt codificate cele două. **Model fit** ne afişează valorile R^2 şi R^2 ajustat. R^2 se numeşte coeficient de determinare şi ne arată cât din variaţia variabilei dependente este explicată de variabilele independente incluse în model. Acesta variază între 0 şi 1, dar noi îl vom transforma în procente pentru că este mai uşor de citit. Un R^2 egal cu 0.56 înseamnă că 56% din variaţia variabilei dependente este explicată de variabilele independente incluse în model. Cu cât este mai mare valoarea, cu atât modelul este mai informativ. Lewis-Beck (1980) enumeră următoarele situaţii pe care trebuie să le avem în vedere când interpretăm valoarea coeficientului de determinare : (a) o valoare mare nu este utilă pentru interpretarea teoretică dacă modelul nu este specificat corect din punct de vedere logic (oferim explicaţii tautologice) ; (b) o valoare mică nu sugerează în mod necesar un model specificat greşit, această situaţie putându-se datora unor relaţii nonliniare între dependentă şi independente.

Pentru că R^2 creşte odată cu introducerea de noi variabile independente în model, consultăm R^2 ajustat, care ia în calcul această situaţie. Sunt situaţii însă, cum ar fi rularea analizei pe eşantioane mici (n sub 100) folosind multe variabile independente (peste 20), când ajustarea poate da greş (Tabachnick şi Fidell, 2007).

În mod uzual, mai bifăm **Confidence intervals**, **R squared change**, **Descriptives**, **Part and partial correlations**, **Collinearity diagnostics**.

Confidence intervals ne oferă intervalele de încredere pentru coeficienţii de regresie nestandardizaţi. Aceştia ne oferă posibilitatea să înţelegem mai realist situaţia explicativă decât estimarea punctuală. Putem vedea limitele între care poate varia valoarea cu care se modifică variabila dependentă atunci când variabila independentă se modifică cu o unitate. Când intervalul este larg, atunci estimarea nu este tocmai utilă din punct de vedere teoretic (Lewis-Beck, 1980).

R squared change este util atunci când utilizăm logica blockurilor. Ne va arăta în ce măsură un nou block de variabile aduce un plus în explicaţia variabilei dependente. Ca şi R^2 ajustat, ia valori între 0 şi 1, dar îl citim în procente pentru o interpretare mai uşoară. Cu cât este mai mare valoarea sa, în condiţiile unei specificări corecte a modelului explicativ, cu atât contribuţia explicativă este mai importantă. Acesta se citeşte împreună cu valoarea nivelului de semnificaţie a testului calculat (*sig. F change*) : atunci când p mai mic sau egal cu 0.05, blockul respectiv de variabile contribuie semnificativ statistic la explicarea variabilei dependente.

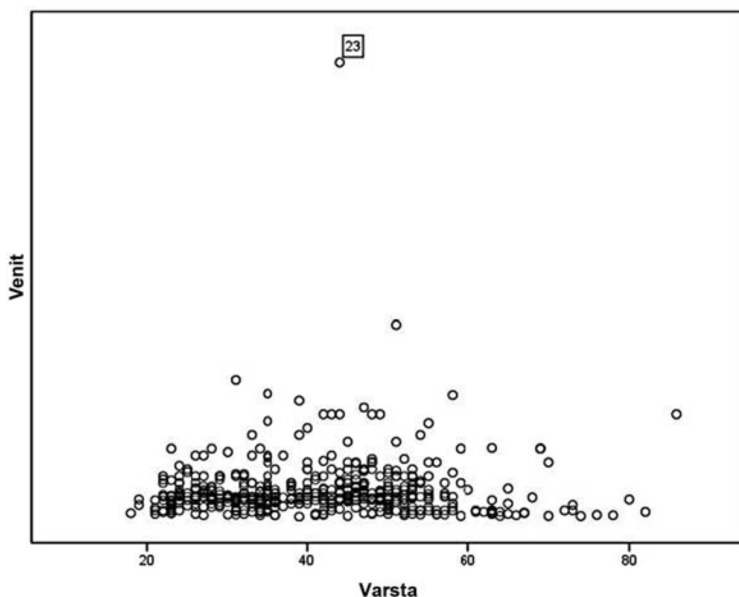
Descriptives calculează mediile şi abaterile standard pentru fiecare variabilă introdusă în ecuaţie şi ne arată volumul eşantionului pentru care sunt efectuate calculele. Putem considera această opţiune ca un punct de control în analiză. Putem calcula media pentru variabilele introduse în model ? Dacă răspunsul este afirmativ, atunci rezultatele analizei de regresie liniare pot fi interpretate. Dacă nu, atunci trebuie să căutăm o soluţie pentru variabila unde nu are sens media. Transformarea în variabile dummy (1/0) este soluţia atunci când trebuie să utilizăm variabile nominale ca variabile independente. Dacă, de exemplu, trebuie să utilizăm religia ca predictor, aceasta având trei categorii, vom alege o categorie de referinţă şi, cu celelalte două, vom realiza două variabile dummy. Pentru alegerea categoriei de referinţă nu există o regulă general valabilă : decizia depinde de interesele analistului. De exemplu, dacă religia are categoriile ortodox, catolic şi protestant, iar interesul cercetătorului este să compare evoluţia dependentei la catolici şi protestanţi prin raportare la ortodocşi, atunci va alege religia ortodoxă ca referinţă şi va crea două dummy-uri astfel :

Variabila iniţială	Variabila dummy 1: catolic	Variabila dummy 2: protestant
Apartenenţa religioasă: 1. ortodox 2. catolic 3. protestant	1 devine 0 <u>2 devine 1</u> 3 devine 0	1 devine 0 2 devine 0 <u>3 devine 1</u>

Este obligatorie introducerea simultană în analiză a celor două variabile dummy. Cele două variabile sunt create folosind meniul **Transform > Recode into different variables**. Media unei variabile dummy indică procentul cazurilor din acea categorie prezente în eşantion şi, dacă eşantionul este reprezentativ pentru o populaţie, respectând structura acesteia, indică procentul cazurilor din acea categorie prezente în populaţie. De exemplu, dacă media variabilei dummy catolic este egală cu 0.31, atunci avem 31% catolici în eşantionul nostru.

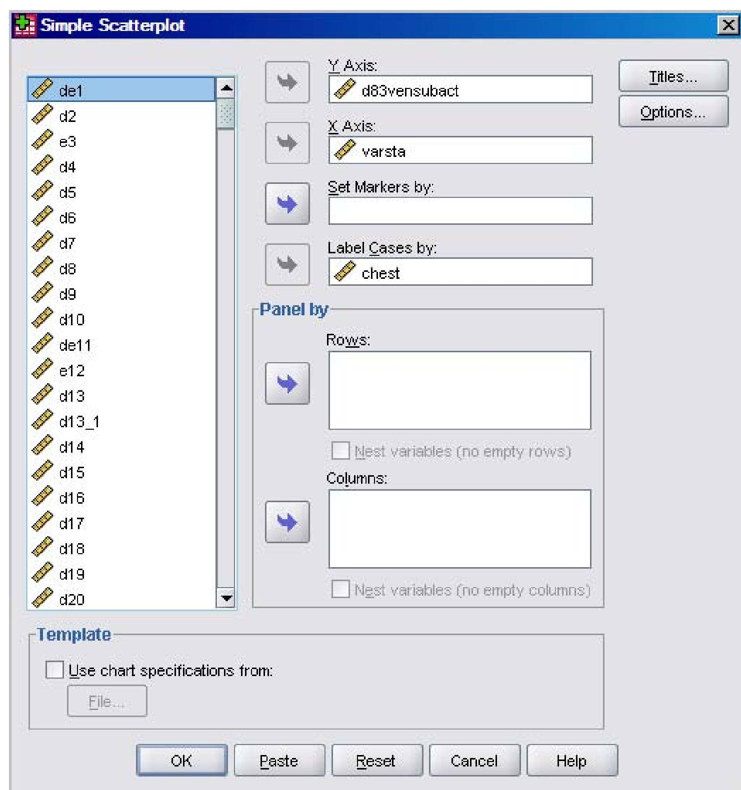
Verificând mediile variabilelor introduse în analiză, ne aducem aminte că acest indicator este influenţat de cazurile extreme. Cazurile extreme sunt univariate sau multivariate. Prezenţa unor astfel de cazuri în analiza de regresie poate modifica serios estimările calculate de program. De aceea, o verificare univariată sau bivariată folosind meniul **Analyze > Descriptives > Explore** este necesară. O altă modalitate complementară, vizuală, constă în realizarea unui grafic scatterplot sau, în limba română, „nor de puncte". În figura 8.5, care prezintă relaţia dintre vârsta măsurată în ani împliniţi şi veniturile persoanelor active pe piaţa muncii din eşantionul DCV ICCV 2003, observăm un caz extrem : o persoană cu vârsta undeva între 40 şi 50 de ani are însumate venituri neaşteptat de mari pentru vârsta sa. Probabil acel venit va ieşi în evidenţă şi la o inspectare univariată a variabilei respective, dar sunt situaţii în care nu se întâmplă aşa, scatterplotul oferind o informaţie foarte utilă în acest sens.

Figura 8.5. Scatterplot care ne arată un caz extrem



Graficul a fost realizat din meniul **Graphs > Legacy dialogs > Scatter – Dot > Simple Scatterplot** (figura 8.6).

Figura 8.6. Meniul Simple Scatterplot



Considerând vârsta variabila explicativă (logic, nici nu am avea cum să o considerăm altfel) şi venitul variabila explicată, cea dintâi este introdusă la **X axis**, iar cea din urmă la **Y axis**. Pentru că, în situaţia în care observăm vreun caz extrem, dorim să îl identificăm uşor, am introdus la **Label Cases by** variabila care conţine id-ul unic al fiecărui respondent care, aici, se numeşte chest. În figura 8.5 eticheta 23 indică id-ul, această valoare putând fi utilizată pentru filtrarea acestui caz din analizele viitoare, de exemplu.

Aceste informaţii sunt utile şi pentru tabelul care conţine corelaţiile bivariate, afişat tot prin alegerea opţiunii **Descriptive statistics**. Corelaţiile bivariate ne ajută să ne facem o primă idee cu privire la relaţiile dintre variabilele incluse în analiză.

Opţiunea **Part and partial correlations** va afişa trei tipuri de corelaţie : **zero-order**, **part** şi **partial**. Dintre acestea ne interesează în mod deosebit corelaţiile semiparţiale pe care SPSS le denumeşte **part correlations**. Această corelaţie, ridicată la pătrat, ne arată contribuţia unică pe care variabila independentă o are la explicarea variabilei dependente. Ne arată cu cât se reduce R^2 dacă acea variabilă independentă este eliminată din ecuaţia de regresie (Tabachnick şi Fidell, 2007). Aceşti autori explică în detaliu diferenţa dintre corelaţia parţială şi cea semiparţială,

atrăgând totodată atenţia că acest mod de interpretare este specific utilizării regresiei multiple standard, adică cea obţinută prin utilizarea metodei **Enter** în SPSS.

În fine, opţiunea **Collinearity Statistics** ne oferă doi indicatori care verifică asumpţia absenţei multicoliniarităţii : **Tolerance** şi **VIF** (**variance inflation factor**). Schroeder, Sjoquist şi Stephan (1986) oferă un exemplu despre ce înseamnă acest lucru : pentru reducerea numărului deceselor rezultate în urma accidentelor auto se introduc simultan două măsuri preventive, purtarea obligatorie a centurii de siguranţă şi pedepsirea aspră a şoferilor prinşi conducând sub influenţa alcoolului. Deşi ambele variabile independente sunt, în esenţă, importante, va fi greu de distins efectul individual al acestora. Aceşti autori atrag atenţia asupra riscului, atunci când există multicoliniaritate, de a întâlni mai des coeficienţi nesemnificativi statistic. Exemplul dat de aceşti autori poate fi completat cu situaţiile în care analistul introduce în analiză variabile independente corelate puternic între ele. Corelaţia puternică poate veni fie din caracterul interşanjabil al indicatorilor (măsoară acelaşi lucru), fie din relaţii de determinare reciprocă. Dacă în exemplul anterior cercetătorul nu poate controla realitatea, interzicând vreuna dintre măsurile preventive, în a doua situaţie, rolul său este de a analiza anterior analizei de regresie atât din punct de vedere logic, cât şi statistic legăturile de determinare dintre variabilele independente. La lista de efecte negative, Field (2009) adaugă şi instabilitatea predicţiei şi limitarea valorilor lui R^2. Revenind la cei doi indicatori, când **Tolerance**, care variază între 0 şi 1, are valori mai mici decât 0.1, asumpţia absenţei multicoliniarităţii este încălcată. **VIF** nu are un interval exact de variaţie. O valoare mai mare decât 10 indică prezenţa multicoliniarităţii (Field, 2009 ; Kline, 2011).

Revenind la exemplul nostru, să vedem ce se întâmplă cu satisfacţia cu viaţa atunci când controlăm veniturile însumate ale persoanelor active pe piaţa muncii, respectiv autopoziţionarea pe scala sărac-bogat. Ipoteza noastră este că ambele au un efect semnificativ statistic pentru că reflectă mecanisme care nu se suprapun perfect. Resursele materiale ne ajută să ne satisfacem nevoile, dar poziţionarea pe scala sărac-bogat implică un proces de comparaţie socială care ne poate face să ne simţim mai săraci (sau mai bogaţi) decât suntem. Depinde care este standardul nostru de referinţă. Înainte de a rula analiza, să apăsăm butonul **Options** (figura 8.4d). Mă opresc asupra secţiunii **Missing Values**, unde este selectat **Exclude cases listwise**. Cunoaştem deja efectele fiecărei metode de tratare a nonrăspunsurilor. Nu o să modificăm nimic în acest meniu.

Outputul analizei este prezentat în continuare. Primul tabel (tabelul 8.4) ne arată media, abaterea standard şi volumul eşantionului pentru care este rulată analiza de regresie. Satisfacţia cu viaţa are media egală cu 5.0 şi abaterea standard egală cu 2.1. Valorile venitului sunt specifice anului 2003, de aici şi modul de prezentare, respectiv sumele mai mici comparativ cu cele de astăzi (atunci când le convertim în lei noi). Observăm că interpretarea mediilor scalelor ordinale, pe care noi le-am considerat de interval, nu este atât de evidentă cum este interpretarea

mediei venitului. Deoarece nonrăspunsurile au fost excluse **listwise**, avem acelaşi volum al eşantionului pe care s-a rulat analiza la toate cele trei variabile (n = 485). Volumul eşantionului a scăzut semnificativ.

Tabelul 8.4. Output regresie liniară multiplă, Descriptives statistics

Descriptive Statistics			
	Mean	Std. Deviation	N
e154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?	5.00	2.146	485
d83vensubact	3973948.45	4341032.795	485
d70 Poziţia pe scala sărăcie-bogăţie	4.43	1.631	485

Tabelul 8.5. Output regresie liniară, Corelaţii bivariate

Correlations				
		e154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENE-RAL?	d83ven-subact	d70 Poziţia pe scala sărăcie-bo-găţie
Pearson Correlation	e154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?	1.000	.251	.662
	d83vensubact	.251	1.000	.204
	d70 Poziţia pe scala sărăcie-bogăţie	.662	.204	1.000
Sig. (1-tailed)	E154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?	.	.000	.000
	D83vensubact	.000	.	.000
	D70 Poziţia pe scala sărăcie-bogăţie	.000	.000	.
N	E154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?	485	485	485
	d83vensubact	485	485	485
	d70 Pozitia pe scala sărăcie-bogăţie	485	485	485

Tabelul de corelaţii din tabelul 8.5 ne oferă o primă imagine a relaţiilor care ne interesează, dar şi informaţii preliminare despre asumpţia absenţei multicoliniarităţii. Este calculat coeficientul Pearson, care variază în intervalul [-1, 1]. Nivelurile de semnificaţie calculate sunt prezentate în rândul **Sig. (1-tailed)**. În principiu, satisfacţia este corelată semnificativ statistic cu ambele variabile. Cei doi indicatori subiectivi, satisfacţia şi autopoziţionarea pe scala sărac-bogat, au o corelaţie mai puternică, lucru aşteptat având în vedere proprietăţile psihometrice similare. Corelaţia mai mică a satisfacţiei cu venitul nu trebuie luată ca atare pentru că ar putea indica prezenţa cazurilor extreme sau chiar a unei relaţii nonlineare.

Tabelul 8.6 prezintă valoarea coeficientului de determinare, **R Square**, şi a coeficientului de determinare ajustat, **Adjusted R Square**. Diferenţele în acest exemplu sunt mici. De regulă, raportăm ambele valori, oferindu-le posibilitatea cititorilor să aprecieze diferenţele. Aşadar, 45% din variaţia satisfacţiei cu viaţa pare să fie explicată de venit şi autopoziţionarea pe scala sărac-bogat. Prima impresie ar fi că am ales bine cele două variabile independente. Testul de semnificaţie este în tabelul ANOVA. Evident, aceasta este o estimare punctuală, de aceea ne putem imagina că ea poate varia în jurul acestei valori.

Tabelul 8.6. Output regresie liniară, R^2

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.673[a]	.452	.450	1.591
a. Predictors: (Constant), d70 Poziţia pe scala sărăcie-bogăţie, d83vensubact				

ANOVA[b]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1008.351	2	504.175	199.085	.000[a]
	Residual	1220.647	482	2.532		
	Total	2228.998	484			
a. Predictors: (Constant), d70 Poziţia pe scala sărăcie-bogăţie, d83vensubact						
b. Dependent Variable: e154 CÂT DE SATISFĂCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?						

Am ajuns la tabelul (tabelul 8.7) care ne oferă informaţiile căutate. Pentru fiecare variabilă independentă ne sunt oferite următoarele informaţii :

- nivelul de semnificaţie (coloana **Sig.**) al testului t care indică dacă între variabila independentă şi variabila dependentă există o relaţie semnificativă statistic. Aici, pentru ambele variabile independente, acesta este mai mic decât pragul 0.05, pe care am decis să îl utilizăm ca referinţă, deci ambele variabile par să influenţeze satisfacţia cu viaţa.

- coeficienţii de regresie nestandardizaţi (coloana **Unstandardized Coefficients – B**). Aceştia ne arată că, atunci când venitul creşte, satisfacţia cu viaţa creşte cu 0.001 puncte pe scală, respectiv că, atunci când individul se consideră mai bogat, satisfacţia cu viaţa creşte cu 0.839 puncte pe scală. Valoarea foarte mică a coeficientului venitului poate însemna : (a) efectul venitului este de fapt mic sau inexistent atunci când controlăm pentru autopoziţionarea pe scala sărac-bogat ; (b) relaţia dintre venit şi satisfacţia cu viaţa nu este liniară, deci ar trebui să revizuim analiza (eliminarea cazurilor extreme dacă există, transformarea variabilelor, introducerea în regresie a pătratului venitului etc.) ; (c) alte asumpţii sunt încălcate.
- coeficienţii de regresie standardizaţi (coloana **Standardized Coefficients – Beta**) sunt folosiţi uneori pentru a spune care dintre predictori are contribuţia cea mai importantă la explicarea variabilei dependente. Totuşi aceştia nu pot fi interpretaţi pentru variabilele dummy (Lewis-Beck, 1980), de aceea ne uităm mai degrabă la pătratul corelaţiilor semiparţiale din coloana **Correlations – Part**.
- indicii care testează absenţa multicoliniarităţii sunt prezentaţi în coloana **Collinearity Statistics**. Indicii de toleranţă sunt foarte mari, având valori peste pragul 0.1, iar **VIF** este mai mic decât 10 pentru ambele independente. Statistic nu există multicoliniaritate. Dar trebuie să ne gândim şi dacă, logic, efectul individual al celor două variabile independente poate fi disociat.

Tabelul 8.7. Output regresie liniară, Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.046	.211		4.952	.000	.631	1.461					
	d83vensubact	.000	.000	.120	3.499	.001	.000	.000	.251	.157	.118	.958	1.043
	d70 Poziţia pe scala sărăcie-bogăţie	.839	.045	.638	18.518	.000	.750	.928	.662	.645	.624	.958	1.043

a. Dependent Variable: e154 CÂT DE SATISFÃCUT SUNTEŢI DE VIAŢA DVS. ÎN GENERAL?

Am înţeles care este logica regresiei liniare multiple şi cum se realizează în SPSS. Pasul următor firesc constă în verificarea tuturor asumpţiilor pe care această analiză le are. Pentru înţelegerea lor vă recomand să parcurgeţi lucrarea scrisă de Berry (1993).

8.3. Exerciţii

Pentru aceste exerciţii utilizăm baza de date şi/sau chestionarul World Values Survey 2012 rezultat(ă/e) în urma aplicării chestionarului în România. Baza de date poate fi descărcată de pe pagina de internet a *Grupului Românesc pentru Studiul Valorilor Sociale* (http : //www.romanianvalues.ro).

1. Citiţi materialul scris de Bogdan Voicu, Horaţiu Rusu şi Mircea Comşa, cu titlul *Atitudini faţă de solidaritate în România*, care a fost publicat în volumul coordonat de Lucian Marina, *Ocupare şi incluziune socială*, apărut la Editura Presa Universitară Clujeană, în 2013.
2. Creaţi variabila dependentă „solidaritate”. Înainte de aceasta, rescalaţi în acord cu modul de lucru al autorilor.
3. Creaţi variabila „orientare de valoare materialistă sau postmaterialistă” în acord cu modul de lucru al autorilor.
4. Inversaţi scala care măsoară importanţa acordată religiei în acord cu modul de lucru al autorilor.
5. Creaţi variabila dummy care măsoară comportamentul religios în acord cu modul de lucru al autorilor.
6. Continuaţi procesul de creare, recodificare, transformare al variabilelor „mândria de a fi român”, „sentimentul apartenenţei naţionale”, „individualism” şi „clasa socială” în acord cu modul de lucru al autorilor.
7. Pregătiţi pentru analiză variabilele „vârstă”, „venit”, „educaţie”, „sex” şi „tip de localitate” în acord cu modul de lucru al autorilor.
8. Rulaţi regresia liniară multiplă în care „solidaritatea” este variabila dependentă, iar toate celelalte sunt independente.
9. Rulaţi din nou regresia liniară multiplă, dar de data aceasta folosiţi blockurile. Ce informaţie suplimentară obţineţi în acest mod ?
10. Realizaţi un raport de două pagini care să descrie rezultatul modelului complet de regresie (cu toţi predictorii) : pe prima pagină este inserat tabelul de regresie, iar pe a doua pagină acesta este comentat cu trimitere la teoriile din textul celor trei autori.

Bibliografie

Agresti, Alan şi Finlay, Barbara. 2008. *Statistical methods for the social sciences*. Upper Saddle River, New Jersey : Prentice Hall International, Inc.

Agresti, Alan şi Franklin, Christine. 2013. *Statistics. The art and science of learning from data*. Boston : Pearson.

Berry, William D. 1993. *Understanding Regression Assumptions*. Newbury Park, CA : Sage.

Bickel, Robert. 2007. *Multilevel Analysis for Applied Research. It's just a regression !* New York : The Guilford Place.

Bradburn, Norman M., Sudman, Seymour şi Wansink, Brian. 2004. *Asking Questions. The Definitive Guide to Questionnaire Design — For Market Research, Political Polls, and Social and Health Questionnaires*, Revised Edition. San Francisco : Jossey-Bass, A Wiley Imprint.

Carifio, James şi Perla Rocco, J. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response format and their antidotes. *Journal of Social Sciences* 3(3) : 106-16.

Carmines, Edward G. şi Zeller, Richard A. 1979. *Reliability and validity assessement*. London : Sage.

Carroll, John B. 1961. The nature of the data, or how to choose a correlation coefficient. *Psychometrika* 26(4).

Chambers, John M., Cleveland, William S., Kleiner, Beat şi Tukey, Paul A. 1983. Graphical methods for data anaylsis. Pacific Groove, California : Wadsworth & Brooks / Cole Publishing Company.

Chelcea, Septimiu. 2007. *Metodologia cercetării sociologice. Metode cantitative şi calitative.* Bucureşti, Editura Economică.

Chen, Peter Y. şi Popovich, Paula M. 2002. *Correlation. Parametric and Nonparametric Measures*. Thousand Oaks, CA : SAGE.

Cramer, Duncan şi Howitt, Dennis. 2004. *The SAGE Dictionary of Statistics. A Practical Resource for Students in The Social Sciences.* London, Thousands Oaks : SAGE.

Cummins, Robert A. 2003. Normative life satisfaction : Measurement issues and a homeostatic model. *Social Indicators Research* 64(2) : 225-56.

de Vaus, David. 2002. *Analyzing social science data. 50 key problems in data analysis*. London : SAGE.

Diener, Ed. 1984. Subjective Well-Being. *Psychological Bulletin* 95(3) : 542-75.

Easterlin, Richard A., Angelescu McVey, Laura, Switek, Małgorzata, Sawangfa Onnicha şi Smith Zweig, Jakueline. 2010. The happiness-income paradox revisited. *Proceedings for the National Academy of Science of the United States of America* 107(52) : 22463-68.

Field, Andy. 2009. *Discovering statistics using SPSS (ans sex and drugs and rock 'n' roll)*. Los Angeles : SAGE.

Fox, John. 1991. *Regression diagnostics.* Newburry Park, CA : SAGE.

Fox, John. 2009. *A mathematical primer for social statistics*. Los Angeles : SAGE.

Gandelman, Nestor, Piani, Giorgina şi Ferre, Zuleika. 2012. Neighborhood Determinants of Quality of Life. *Journal of Happiness Studies* 13 : 547-63.

Good, Phillip I. şi Hardin, James W. 2012. *Common errors in statistics (and how to avoid them)*. Hoboken, New Jersey : John Wiley and Sons, Inc.

Graham, Carol şi Pettinato, Stefano. 2006. Frustrated achievers : winners, losers, and subjective well-being in Peru's emerging economy. *The ANNALS of the American Academy of Political and Social Science* 606 : 128-53.

Hagerty, Michael R. şi Veenhoven, Ruut. 2003. Wealth and happiness revisited. Growing wealth of nations does go with greater happiness. *Social Indicators Research* 64 : 1-27.

Hair, Joseph F., Black, William C., Babin, Barry J. şi Anderson, Rolph E. 2010. *Multivariate data analysis. A global perspective.* Upper Saddle River N.J. : Pearson.

Hartwig, Frederick şi Dearing, Brian E. 1979. *Exploratory data analysis.* Newbury Park, CA : SAGE.

Henkel, Ramon E. 1976. *Tests of significance*. Newbury Park, CA : SAGE.

Hooghe, Marc şi Vanhoutte, Bram. 2011. Subjective Well-Being and Social Capital in Belgian Communities. The Impact of Community Characteristics on Subjective Well-Being Indicators in Belgium. *Social Indicators Research* 100(1) : 17-36.

Inglehart, Ronald, Foa, Roberto, Peterson, Christopher şi Welzel, Christian. 2008. Development, freedom, rising happiness. A global perspective, 1981–2007. *Perspectives on Psychological Science* 264-285(3).

Inglehart, Ronald şi Welzel, Christian. 2005. *Cultural Change and Democracy : The Human Development Sequence*. New York and Cambridge : Cambridge University Press.

Iversen, Gudmund R. şi Norpoth, Helmut. 1987. *Analysis of variance.* Thousand Oaks, CA : SAGE.

Jaccard, James şi Jacoby, Jacoby. 2010. *Theory Construction and Model-Building Skills. A Practical Guide for Social Scientists.* New York : The Guilford Press.

Jacoby, William G. 1997. *Statistical graphics for univariate and bivariate data*. Newbury Park, CA : Sage.

Jacoby, William G. 1998. *Statistical graphics for visualizing multivariate data*. Newbury Park, CA : Sage.

Kline, Rex B. 2004. *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC : American Psychological Association.

Kline, Rex B. 2011. *Principles and Practice of Structural Equation Modeling*. New York : The Guilford Press.

Lelkes, Orsolya. 2008. *Happiness Across the Life Cycle : Exploring Age-Specific Preferences*. Policy Brief (2).

Levy, Paul S. şi Lemeshow, Stanley. 2008. *Sampling of populations. Methods and applications*. Hoboken, New Jersey : John Wiley & Sons, Inc.

Lewis-Beck, Michael S. 1980. Applied Regression : An Introduction. Newburry Park, CA : SAGE.

Likert, Rensis. 1932. *A technique for the measurement of attitudes*. Archives of Psychology 22(140) : 5-55.

Malhotra, Naresh K. şi Birks, David F. 2007. *Marketing Research. An Applied Approach*. Harlow, England : Prentice Hall Financial Times.

Mărginean, Ioan. 1982. *Măsurarea în sociologie*. Bucureşti : Editura Ştiinţifică şi Enciclopedică.

Mărginean, Ioan. 2005. *Semnificaţia cercetărilor de calitate a vieţii*. Pp. 25-60 în *Calitatea vieţii în România*, editată de Ioan Mărginean, Ana Bălaşa şi Cătălin Zamfir. Bucureşti : Editura Expert.

Michalos, Alex C. 2005. *Multiple discrepancies theory (MDT) (1985)*. Pp. 305-72 in *Citation Classics from Social Indicators Research. The Most Cited Articles Edited and Introduced by Alex C. Michalos*, *edited by Alex C. Michalos*. Dordrecht, The Netherlands : Springer.

Mikucka, Małgorzata. 2012. *Individualist culture lowers well-being of the unemployed due to weaker family support norm. Evidence for Europe.* CEPS/Instead, Luxembourg.

Mohr, Lawrence B. 1990. Understanding Significance Testing. Newburry Park, CA : SAGE.

Raudenbush, Stephen W., Bryk, Anthony S., Cheong, Yuk Fail, Congdon, Richard T. Jr. şi du Thoit, Mathilda. 2011. HLM 7. *Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL. : SSI Scientific Software International.

Reynolds, H.T. 1984. *Analysis of nominal data*. Newbury Park, CA : Sage.

Rotariu, Traian (coord.), Bădescu, Gabriel, Culic, Irina, Mezei, Elemer şi Mureşan, Cornelia. 2006. *Metode statistice aplicate în ştiinţele sociale*. Iaşi : Polirom.

Rughiniş, Cosima. 2007. *Explicaţia sociologică*. Iaşi : Polirom.

Saris, Willem E. şi Gallhofer, Irmtraud N. 2007. *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, New Jersey : John Wiley & Sons, Inc.

Schroeder, Larry D., Sioquist, David L. şi Stephen, Paula E. 1986. *Understanding regression analysis. An introductory guide*. Newbury Park, CA : SAGE.

Schwartz, Barry. 2004. *The paradox of choice. Why more is less*. New York : Harper Collins.

Stoop, Ineke A.L. 2005. *The Hunt for the Last Respondent. Nonresponse in sample surveys.* Social and Cultural Planning Office for Netherlands : The Hague.

Tabachnick, Barbara G. şi Fidell, Linda S. 2007. *Using Multivariate Statistics*. Boston : Pearson.

Tufte, Edward R. 2001. *The visual display of quantitative information.* Cheshire, Connecticut : Graphics Press.

Tversky, Amos şi Kahneman, Daniel. 1991. *Loss Aversion in Riskless Choice : A Reference-Dependent Model. The Quarterly Journal of Economics 106* (4 [Nov.]) : 1039-61.

Veenhoven, Ruut. 1996. The study of life satisfaction. Pp. 11-48 in *A comparative study of satisfaction with life in Europe*, editată de Saris, W.E., Veenhoven, R., Scherpenzeel, A.C. şi Bunting, B : Eötvös University Press.

Voicu, Bogdan, Rusu, Horaţiu şi Comşa, Mircea. 2013. Atitudini faţă de solidaritate în România. Pp. 17-44 în *Ocupare şi incluziune socială*, Lucian Marina. Cluj : Presa Universitară Clujeană.

Zamfir, Cătălin (coordonator), Popescu, Ion-Andrei, Ştefănescu, Ştefan, Teodorescu, Alin, Vlăsceanu, Lazăr şi Zamfir, Elena. 1984. *Indicatori şi surse de variaţie a calităţii vieţii*. Bucureşti : Editura Academiei Republicii Socialiste România.

Zimmerman, Anke C. şi Easterlin, Richard A. 2006. Happily Ever After ? Cohabitation, Marriage, Divorce, and Happiness in Germany. *Population and Development Review* 32(3) : 511-28.

COLLEGIUM
Sociologie. Antropologie

au mai apărut:

Petru Iluţ – *Sinele şi cunoaşterea lui. Teme actuale de psihosociologie*
Marie-Odile Géraud, Olivier Leservoisier, Richard Pottier – *Noţiunile-cheie ale etnologiei. Analize şi texte*
Marian Preda – *Politica socială românească între sărăcie şi globalizare*
Emile Durkheim – *Regulile metodei sociologice*
Albert Ogien – *Sociologia devianţei*
Traian Rotariu – *Demografie şi sociologia populaţiei. Fenomene demografice*
Ioan Mihăilescu – *Sociologie generală. Concepte fundamentale şi studii de caz*
W. Richard Scott – *Instituţii şi organizaţii*
Irina Culic – *Metode avansate în cercetarea socială. Analiza multivariată de interdependenţă*
Petru Iluţ – *Valori, atitudini şi comportamente sociale. Teme actuale de psihosociologie*
Cătălin Zamfir – *O analiză critică a tranziţiei. Ce va fi „după”*
Gilles Ferréol, Guy Jucquois (coord.) – *Dicţionarul alterităţii şi al relaţiilor interculturale*
Robert K. Yin – *Studiul de caz. Designul, analiza şi colectarea datelor*
Richard A. Krueger, Mary Anne Casey – *Metoda focus grup. Ghid practic pentru cercetarea aplicată*
Ronald F. King – *Strategia cercetării. Treisprezece cursuri despre elementele ştiinţelor sociale*
Petru Iluţ – *Sociopsihologia şi antropologia familiei*
Dumitru Sandu – *Dezvoltare comunitară. Cercetare, practică, ideologie*
Cătălin Zamfir – *Spre o paradigmă a gândirii sociologice*
Mircea Agabrian – *Analiza de conţinut*
Adrian Hatos – *Sociologia educaţiei*
Cătălin Zamfir, Laura Stoica (coord.) – *O nouă provocare : dezvoltarea socială*
Marian Preda – *Comportament organizaţional*
Mihai Păunescu – *Organizare şi câmpuri organizaţionale. O analiză instituţională*
Robert Atkinson – *Povestea vieţii. Interviul*
Traian Rotariu, Petru Iluţ – *Ancheta sociologică şi sondajul de opinie. Teorie şi practică* (ediţia a II-a)
Amia Lieblich, Rivka Tuval-Mashiach, Tamar Zilber – *Cercetarea narativă. Citire, analiză şi interpretare*
Lazăr Vlăsceanu – *Sociologie şi modernitate. Tranziţii spre modernitatea reflexivă*
Jacques Coenen-Huther – *Sociologia elitelor*
Cosima Rughiniş – *Explicaţia sociologică*
Dumitru Sandu (coord.), Cosmin Câmpean, Lucian Marina, Mihaela Peter, Vasile Şoflău – *Practica dezvoltării comunitare*
Cătălin Zamfir, Simona Stănescu (coord.) – *Enciclopedia dezvoltării sociale*
Mihai Coman – *Introducere în antropologia culturală. Mitul şi ritul*
Traian Rotariu – *Demografie şi sociologia populaţiei. Structuri şi procese demografice*
Vintilă Mihăilescu – *Antropologie. Cinci introduceri* (ediţia a II-a)
Remus Gabriel Anghel, István Horváth (coord.) – *Sociologia migraţiei. Teorii şi studii de caz româneşti*
Nicu Gavriluţă – *Antropologie socială şi culturală*
Petru Iluţ – *Psihologie socială şi sociopsihologie. Teme recurente şi noi viziuni*
Raluca Popescu – *Introducere în sociologia familiei. Familia românească în societatea contemporană*
Călin Cotoi – *Introducere în antropologia politică*
Marian Preda (coord.) – *Riscuri şi inechităţi sociale în România. Raportul Comisiei Prezidenţiale pentru Analiza Riscurilor Sociale şi Demografice*
Vintilă Mihăilescu (coord.) – *Etnografii urbane. Cotidianul văzut de aproape*
Earl Babbie – *Practica cercetării sociale*
Victor Jupp (coord.) – *Dicţionar al metodelor de cercetare socială*

Cristina Gavriluţă, Nicu Gavriluţă – *Sociologia sportului. Teorie, metode, aplicaţii*
Dumitru Sandu – *Lumile sociale ale migraţiei româneşti în străinătate*
Traian Rotariu – *Studii demografice*
Mihaela Vlăsceanu – *Economie socială şi antreprenoriat. O analiză a sectorului nonprofit*
Laura Grünberg (coord.) – *Introducere în sociologia corpului. Teme, perspective şi experienţe întrupate*
Lazăr Vlăsceanu (coord.) – *Sociologie*
Alex Preda – *Introducere în sociologia pieţelor. Informaţie, cunoaştere şi viaţă economică*
Martine Segalen – *Sociologia familiei*
Camelia Beciu – *Sociologia comunicării şi a spaţiului public. Concepte, teme, analize*
Mihai Dinu Gheorghiu, Monique de Saint Martin (coord.) în colaborare cu Bénédicte de Montvalon – *Educaţie şi frontiere sociale : Franţa, România, Brazilia, Suedia*
Gabriel Jderu – *Introducere în sociologia emoţiilor*
Traian Rotariu, Vergil Voineagu (coord.) – *Inerţie şi schimbare. Dimensiuni sociale ale tranziţiei în România*
Cosima Rughiniş – *Măsurarea sociologică. Teorii şi practici ale cuantificării*
Lazăr Vlăsceanu – *Introducere în metodologia cercetării sociologice*
Petru Iluţ (coord.) – *În căutare de principii. Epistemologie şi metodologie socială aplicată*
Nicu Gavriluţă – *Sociologia religiilor. Credinţe, ritualuri, ideologii*
Marian-Gabriel Hâncean – *Reţelele sociale. Teorie, metodologie şi aplicaţii*
Marian Vasile – *Introducere în SPSS pentru cercetarea socială şi de piaţă. O perspectivă aplicată*

www.polirom.ro

Redactor: Dan Mironescu
Coperta: Radu Răileanu
Tehnoredactor: Radu Căpraru

Bun de tipar: februarie 2014. Apărut: 2014
Editura Polirom, B-dul Carol I nr. 4 • P.O. BOX 266
700506, Iaşi, Tel. & Fax: (0232) 21.41.00; (0232) 21.41.11;
(0232) 21.74.40 (difuzare); E-mail: office@polirom.ro
Bucureşti, Splaiul Unirii nr. 6, bl. B3A, sc. 1, et. 1,
sector 4, 040031, O.P. 53 • C.P. 15-728
Tel.: (021) 313.89.78; E-mail: office.bucuresti@polirom.ro

Tiparul executat la S.C. Tipo-Lidana S.R.L.
Calea Unirii nr. 35, Suceava
Tel.: 0230/517.518; Fax: 0330/401.062
E-mail: office@tipolidana.ro; www.tipolidana.ro